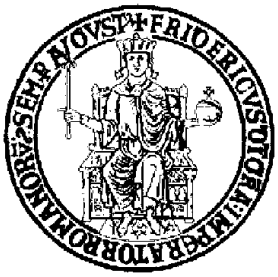
UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II



FACOLTÀ DI SCIENZE POLITICHE

DIPARTIMENTO DI TEORIE E METODI DELLE SCIENZE UMANE E SOCIALI

SCUOLA DI DOTTORATO IN
SCIENZE PSICOLOGICHE, PEDAGOGICHE E LINGUISTICHE
DOTTORATO DI RICERCA

IN

LINGUA INGLESE PER SCOPI SPECIALI

XXIV CICLO

TESI DI DOTTORATO

Collocations in University degree descriptions: an evaluation of lexical association measures

CANDIDATO
Dott. Adriano Ferraresi

RELATORE

Prof.ssa Silvia Bernardini

COORDINATORE

Prof.ssa Gabriella Di Martino

NAPOLI 2011

Abstract

This thesis investigates the notion of collocation and the effects of lexical association measures triangulating corpus, lexicographic and experimental evidence. Focusing on institutional academic English, and in particular on the genre of degree course descriptions, a special-purpose corpus is constructed semi-automatically from the web. Two widely-used lexical association measures (Mutual Information and Log-likelihood), a relatively recent one (Lexical Gravity), and bare cooccurrence frequency are used to extract collocation candidates from this corpus using a stratified sampling technique. The 99 phrases thus selected are searched for in two dictionaries (a collocation dictionary and a general purpose learner dictionary), presented to expert informants who evaluate their acceptability, and used in a lexical decision task.

The results of these evaluation tasks suggest that *a)* none of the measures significantly outperforms the others in extracting salient word pairs, even though bare frequency seems to perform marginally better than the others in the lexical decision task, and MI in the acceptability judgement task; *b)* different measures target different types of phrases (both in terms of the distinction between free/restricted combinations, and in terms of their degree of specialization) ; *c)* some measures perform better in the top range (e.g. Lexical Gravity), while for other measures the best results are scattered in different frequency ranges (e.g. Mutual Information); *d)* native speaker and non native speaker expert informants seem to evaluate collocativity in similar ways, even though non natives are more conservative, giving less extreme scores; *e)* the acceptability judgement questionnaire and the lexical decision task, performed on different groups using different experimental methodologies, provide converging evidence: the expressions that experts find to be the most acceptable are also recognized faster and more accurately by subjects in the test. In turn, *f)* this experimental evidence is correlated with the corpus evidence extracted by the association measures.

The implications of these findings are manyfold. On the theoretical side, they confirm that corpora and lexical association measures provide evidence that is coherent with that obtained from experimental methods targeting language competence. On the descriptive side, the study suggests that the phraseology of degree course description is characterized by a mix of disciplinary terms ("cochlear implants", "linear algebra") and core phraseology typical of the genre ("wide genre", "open days"), as well as showing which lexical association measures are more appropriate for targeting the former (Mutual Information) or the latter (Frequency or Lexical Gravity). Lastly on the applied side, these findings can be used to provide guidelines as to the best lexical association measures to use depending on the type of phrases one wants to extract, the amount of manual filtering that can be applied to the task, the number of phrases to extract and so forth.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The object of this thesis is a comparative evaluation of different statistical measures for the automatic extraction of lexical collocations from an ESP corpus, using lexicographic, informant and psycholinguistic evidence. There are three related aspects to this theme. The first is mainly theoretical, and concerns the overlap between a performance-based view of collocation (i.e. collocations as retrieved from a corpus, regardless of the method employed for retrieval) and its competence-based counterpart (relying on implicit evidence of psychological salience or explicit endorsement of collocation status). The second, more practical/methodological concern has to do with comparing the different results yielded by three different measures of collocationality and a baseline, in order to determine which measure better matches the competence-based evidence collected. The third, descriptive issue addressed in this thesis regards the typical phraseology employed within a well-defined genre, namely degree course descriptions published by British universities on the web.

A (very) loose corpus-based definition might describe collocations as sequences of words that occur repeatedly in texts, and that do so because they are "the preferred way of putting things" (Kennedy 1992): for instance, based on the relative frequency of occurrence of "final year" and "concluding year" in a corpus of degree course descriptions, it is possible to conclude that students who are at the end of their studies are more likely to be described as being in the former than in the latter, regardless of the fact that the two adjectives are near-perfect synonyms in context. The terms "preferred" and "likely" in the previous sentence hint at the fact that repeated cooccurrence is hypothesized not to be a random feature of texts, but rather the textual instantiation of psychological salience: collocations form a crucial part of a speakers' mental lexicon, *therefore* they are uttered or written often, *therefore* they are highly frequent in texts. While this seems a fair assumption, that is taken for granted, either implicitly or explictly, in most studies on the topic, the actual relation between a performance- and a

competence-oriented view of collocation, i.e. between corpus and psycholinguistic data, is still underexplored in the field of corpus linguistics (Gilquin and Gries 2009).

Moving on to the second issue, different statistical measures have been proposed and are currently used in the literature, that give more prominence (i.e. a higher collocativity score and rank) to one or another sequence (say, "second year" vs. "final year"). The question then arises as to what is the "best" measure of collocativity available, or what measure is able to retrieve from corpora the highest number of salient collocations while minimizing or scoring down non-collocations. Evidence that "final year" (but not "second year") is implicitly or explicitly recognized as a collocation by speakers of English would suggest that it is memorized as a single unit, i.e. that it is part of their mental lexicon rather than being compositional. An AM that gives a higher score to "final year" than to "second year" better reflects human intuition than one that does the opposite, and there are obvious descriptive/theoretical and practical/methodological advantages in knowing which does what.

Finally, the collocations evaluated in this thesis are extracted from a purpose-built corpus of BA degree course descriptions collected through a semi-automatic procedure from the websites of British universities. This genre was selected since it provides a well-defined and clearly recognizable subset of an ESP variety that is currently the object of both descriptive (Biber 2006) and applied interest (Depraetere et al. 2011). While the analysis is limited to adjective-noun pairs, it does provide insights about typical phrases used in this native variety of English, that are of interested both on their own and for subsequent comparisons with lingua franca varieties.

The thesis is structured as follows. Chapter 2 describes the theoretical background to the thesis, i.e. the ESP under study (institutional academic English, and in particular degree course descriptions) and the aspects of collocation studies of more immediate relevance to the present concerns, namely frequency-oriented views, statistical methods and process-oriented perspectives. Chapter 3 focuses on methodological aspects, providing a detailed account of the various phases of the research - formulation of the research hypotheses, set up of the corpus, evaluation tasks, and statistical methods used in the analysis of results. Chapter 4 reports on the results obtained in the three evaluation tasks, carrying out extensive quantitative and qualitative comparisons and discussing points of contact and differences observed. Finally Chapter 5 recaps on the main findings of the thesis, comments on their theoretical, methodological and applied relevance, and makes suggestions for further work.

Chapter 2

Background: Research on institutional academic English and on collocation

2.1 Overview of the Chapter

This chapter presents the two-sided background to the present thesis, i.e. research on institutional academic English on the one hand, and research on collocations on the other. As concerns the first aspect, Section 2.2.1 surveys literature on this LSP, that has received limited attention so far despite its being of descriptive interest and practical importance, while Section 2.2.2 looks at the specific genre focused upon in the thesis, namely degree course descriptions, focusing on criteria of genrehood that are applicable to corpus construction. The second part of the Chapter reviews previous work on collocation (2.3.1), presents relevant parameters proposed for the identification and categorization of collocations (2.3.2), and discusses the two aspects of collocation studies that are of immediate relevance for the purposes of this work, namely statistical (2.3.3) and experimental (2.3.4) methods.

2.2 Institutional academic English and degree course descriptions

2.2.1 Why institutional academic English

Within English for Special Purposes and English for Academic Purposes (henceforth ESP and EAP respectively) substantial work has been devoted to academic research genres, i.e. to the discourse used within academia for knowledge sharing

(Ph.D. dissertations and defences, research articles and talks, as well as sub-genres such as article abstracts and introductions; see e.g. Swales (2004) and the references in Gesuato (2011)). Recent years have also witnessed a surge of interest in genres which are arguably more marginal in terms of scientific achievement, but equally central to academic life, e.g. book reviews (Römer 2010), grant proposals (Connor and Upton 2004), thesis acknowledgements, doctoral prize applications and bio statements (Hyland 2011).

Such genres situate themselves midway between the strictly disciplinary genres traditionally focused upon in discourse and genre studies (e.g. the research article), and the genres used for everyday institutional academic communication – especially between institutions and their (prospective or current) students – i.e. syllabi, course packs, welcome messages, mission statements, announcements and so forth. Due to their subservient "managing" function with respect to the research genres, the latter have so far been largely neglected as an object of study, with some notable exceptions that will be discussed here. Yet this state of affairs is bound to change, as Universities worldwide place more and more importance on strategies for effectively managing relations with prospective and current students and alumni.

Landmark works focusing on institutional academic genres have been produced mainly within applied corpus linguistics and critical discourse analysis. The former are motivated by the observation that international students wishing to study in English-medium institutions need to understand many types of texts, including complex hybrid ones in which informative, directive, and promotional functions often coexist (Gesuato 2011). The latter spring from concerns with the increasing tendency for Universities to adopt business models and transform education into a saleable good, which are hypothesised to be reflected in their discursive practices.

Within the EAP/ESP approach, Biber (2006) provides a full-fledged account of the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) corpus funded by the U.S. Educational Testing Service, which includes both academic and institutional/management genres. Understanding the latter (e.g. handbooks, catalogues, programme web pages, course syllabi) is suggested to be of great importance for international students wishing to study in English-medium Universities (on the course syllabus in particular see also Afros and Schryer (2009)). The volume focuses on vocabulary use, stance expressions, grammatical and register variation and phraseology, specifically lexical bundles. While the latter are not to be confused with collocations, they are certainly similar in terms of the patterns they focus upon (recurrent non-idiomatic phrases). The comparison of academic and non-academic registers in terms of presence of lexical bundles suggests that they are much more frequent in the latter than in the former, with institutional texts preferring referential expressions, and management texts pre-

ferring stance expressions. The relevance of non-disciplinary genres for applied linguistics purposes is also endorsed by the builders of the well-known MICASE corpus, constructed at the University of Michigan, which includes samples of spoken academic registers not limited to lectures and seminars but also including more informal, everyday interaction on campus such as service encounters and campus tours (Simpson-Vlach and Leicher 2006). These large-scale research efforts seem to point to an increasing attention devoted to the communicative strategies used for effective communication in institutional academic English by the applied and corpus linguistics communities.

Looking at the discursive practices of tertiary education institutions from the perspective of Critical Discourse Analysis, Fairclough (1993:143) suggested that these were "in the process of being transformed through the increasing salience within higher education of promotion as a communicative function" and wondered "what is happening to [...] authority relations between academics and students, academic institutions and the public, etc.?". More recently Swales (2004:9) surveys scene-setting trends in academic communication pointing out how the commodification of education "has been accompanied by language that emphasizes "reenvisioning", "remissioning", or "reengineering", and by a shift in curricular perspective to the needs of the students (now seen as "customers") as opposed to the scholarly expectations of a discipline or the traditional offerings of a department". Analysing a corpus of nonprofit sector texts (including mission statements and deans' welcoming addresses), Mautner (2005:38) shows how Universities borrow commercial models, using persuasive style and "for-profit" language, in particular "[l]exical imports from the business domain".

In the globalized higher-education market, institutions based in countries from the expanding circle (Kachru 1985) are also under increasing pressure to master institutional academic genres in English. This is especially true of Europe at the moment. To achieve the strategic objectives of the Bologna Process, Universities are required to recruit international students, as well as to attract exchange staff and students through mobility programmes. For this internationalisation process to be successful, availability of courses and information in English is essential (Altbach and Knight 2007). Yet, if one takes Italy as a case in point, and quickly browses the web pages of Italian Universities, it is clear that this requirement has been implemented to a limited degree. Interventions aimed at supporting multilingual communicative strategies in the institutional/administrative domain are therefore needed – and strongly encouraged by the EU, see e.g. Depraetere et al. (2011) – to strengthen internationalisation policies, and these in turn require descriptive insights obtained from corpus-based studies in which native and translated or *lingua franca* texts are compared.

The crucial importance of English as a *lingua franca*, especially in scientific and academic international settings, is nowadays widely recognised and has stim-

ulated a number of comparative studies (some corpus-based) analysing non-native varieties against the background of standard "native" varieties Seidlhofer (2001), Mauranen (2003), Jenkins (2007). To the best of my knowledge, however, no in-depth studies to date have been devoted to the discursive features of institutional English as it is used on the websites of European Universities, nor has this *lingua franca* variety of English been compared to native varieties within the EU context (Bernardini et al. (2010) is a first step in this direction). Methodologically-oriented studies of the native benchmark such as the one presented in this thesis would therefore also be essential to pave the way for contrastive analyses of *lingua franca* varieties.

2.2.2 The genre under study: Degree course descriptions

2.2.2.1 Why this genre

Degree course descriptions occupy a central position among institutional academic genres. At the descriptive level, they represent typical examples of texts included by Biber (2006) in the category of "institutional writing", i.e. the type of written material that is "required reading for the prospective students attempting to navigate the maze of university requirements and services" (Biber 2006:26). In practical terms, these texts also play a crucial role in Universities' internationalisation efforts, insofar as they are likely to be a primary source of information for students deciding which degree course to attend. Their strategic importance as a genre is also acknowledged by an ongoing EU-funded project which aims at providing machine translation tools specifically tailored for translating course syllabi and degree programme descriptions between English and 8 other languages (Depraetere et al. 2011).

2.2.2.2 Internal and external criteria of genrehood

In this section a brief description of the genre is provided. While very many criteria can be applied to the description of a genre (cf. the discussion in Swales (1990:Part I), here the main purpose is to provide an account of those that can be used for selecting appropriate specimen for corpus inclusion. We will therefore use as a frame of reference the preliminary recommendations on text typology (Sinclair and Ball 1996) produced as part of the influential EAGLES guidelines on language engineering standards. The EAGLES typology distinguishes between "external" and "internal" text typology criteria, i.e., broadly speaking, between extralinguistic features ("features of the nonlinguistic environment [...] in which the texts occurred", (Sinclair and Ball 1996:unpaginated)) and linguistic features (topic, aspects of the style etc. of a text). While the inextricable interdependence of these two dimensions is explicitly acknowledged, the former is suggested to be

the primary dimension according to which a text should be described (cf. also Sinclair (2004)). Therefore we will focus specifically on "text external" criteria. As will be made clearer in Section 3.3, such characterization is particularly relevant when (semi-)automated methods are used for corpus building. These methods usually rely on searching the web for specific words that are deemed to be relevant for the domain under consideration (Baroni and Bernardini 2004), i.e. they rely on "text internal" criteria. According to Sinclair and Ball (1996), this is undesirable, since "classification of texts based [...] on internal criteria does not give prominence to the sociological environment of the text, thus obscuring the relationship between the linguistic and non-linguistic criteria". Establishing a set of external, genre-defining criteria to be implemented during the corpus construction phase is thus a desirable preliminary step if one is to target specific texts without relying on internal criteria only (i.e. the words they contain). These criteria are presented here (rather than in 3.3) because they can also give the reader a clearer idea of the textual population targeted in our study.

2.2.2.3 Criteria for genre/corpus inclusion

Three main external criteria are proposed in the EAGLES guidelines, and each of them is further subdivided into subcategories:[1] a. origin of the text (i.e. the people involved in the process of text creation, including subcategories like author and publisher); b. state (i.e. the mode of transmission of the text, whether written or spoken and the medium of its publication, e.g. a printed book, a newspaper, an electronic publication); and finally c. aims (i.e. the intended audience of the text and its communicative purpose). Evidence for classifying texts according to these external criteria can either be "circumstantial" (i.e., coming from outside the text) or "reflexive" (i.e., based on statements within the text). In the following subsections, degree course descriptions are characterized in terms of these parameters, using both circumstantial and reflexive evidence.

2.2.2.3.1 Origin As is the case with most institutional texts (Drew and Heritage 1992), online degree course descriptions never mention explicitly their author. It can be hypothesized that multiple authors are behind these texts, i.e. both experts in the discipline(s) making the object of the degree programme (in all likelihood faculty academic staff), and increasingly also professional editors and communications consultants (Mautner 2005:34). On the contrary, the publisher, or originator (Sinclair 2004), of the texts is known, and this should be seen as the most relevant criterion to describe the texts' origin. The University publishing

[1]The original typology includes a wider variety of sub-categories (e.g. the and sex of the writers, the text's copyright holder, etc.). I am including here only those which are relevant for the characterization of the genre under consideration.

the text (and offering the degree) is present on each page thanks to contextual graphics material such as logos, links and images, as well as through the page title and URL. A brief aside is in order, to consider the question whether the authors/originators of degree course descriptions actually form a homogeneous discourse community or not, this being one of the defining criterion for assigning "genrehood" according to Swales (1990) and Bhatia (1993). Discourse communities have tended to be defined in terms of their common academic or scientific interests (cf. features such as the use of shared terminology and content expertise (Swales 1987)). Yet this might equally be a consequence of the fact that more attention has traditionally been placed in genre studies on academic/scientific genres than on institutional ones. While the question cannot be settled here, in terms of communality of interests, mechanisms for intercommunication, provision of information and feedback, and especially development of shared discoursal expectations (Swales 1987), academics and staff based in different Universities within a single country seem likely to share substantial common ground, and thus form a discourse community, though not a disciplinary one.

2.2.2.3.2 State In this work the web version of degree course descriptions is considered. Usually a paper form of the same documents is provided by Universities (cf. Afros and Schryer (2009)), but the Web version is likely to have more global reach among students, as well as being less costly to produce and maintain, for which reasons it is nowadays favoured by institutions (who also, however, make available printer-friendly versions). Far from being mere reproductions of texts published in traditional prospectuses, the web pages of UK degree course descriptions tend to make full use of the medium. For instance, they may use the left and/or right columns to provide quick facts, contact information or useful links; several pages are split into "tabs", so as to make each part short enough not to require scrolling down the page; menus allow moving between e.g. content outline, requirements, admission tests, job prospects of a single course, among courses within a single department, or among departments. Graphics (tables, graphs, logos), photos, sound and video (e.g. of interviews with current students) provide a rich contextual apparatus complementing the actual text.

2.2.2.3.3 Aim This criterion is central to the characterization of this genre. The intended audience primarily consists of prospective students looking to decide what degree course to choose among the many on offer. As a result, the aim is both informative and promotional (Caiazzo 2010). Focusing on reflexive evidence, the informative function is signalled by section titles such as "key facts", "entry requirements", "contact details", "how to apply", "further information", which clearly suggest that the corresponding texts give practical information about the

courses. The promotional aspect is instead evident in links and contextual menus addressing the reader and making ample use of imperative forms of verbs (e.g. "Study here", "Visit/contact us", "Find out more"). Multimodal contents are also very often used promotionally – i.e, to suggest a relaxed, welcoming environment conducive to learning as well as a rich social life and/or social inclusion, depending on institutional priorities. This is in line with the tendency for public sector organisations to become "purposefully multimodal, with pictures and graphic elements taking on an ever more salient role in message design" (Mautner 2005:37).

2.2.2.4 A closing note on internal criteria

According to the EAGLES guidelines, "two central parameters of the classification of texts are better described using internal, or text-linguistic, rather than external, or sociocultural, criteria" (Sinclair and Ball 1996:unpaginated). These are text "topic" and text "style". Since internal criteria should not be employed when delimiting a population for purposes of corpus construction (2.2.2.2), these parameters are not discussed here. Yet for the purposes of this work it is interesting to note in passing the central role played by collocations for the analysis of these defining aspects of genrehood:

> [T]he clustering of collocates [...] gives a more accurate identification of the topic of the text than simple keywords. In style, types of word combination are clues to style types. [Collocation can also be used] to classify genres, showing that the same word is characteristically associated with certain collocates in particular types of writing and speaking. (Sinclair and Ball 1996:unpaginated)

2.3 Collocation: an overview

Even though the notion of collocation dates back at least to the first half of the 20th century (see the discussion in Sinclair et al. (1970)), it has enjoyed considerable popularity in the last two decades, following a shift of focus away from syntax and rule-based approaches and toward the lexicon and usage-based approaches, that has characterized both theoretical (Croft and Cruse 2004; Fillmore et al. 1988) and applied linguistics (Lewis 1993; O'Dell and McCarthy 2008; Willis 2001). The widespread use of corpus-based methods of language analysis has provided a wealth of descriptive insights confirming the central role played by the idiom principle (Sinclair 1991) in language use. More recently, researchers have started to investigate the relationship between descriptive insights about collocations obtained from product-oriented studies conducted on corpora, and

evidence about communicative competence obtained from process-oriented psycholinguistic studies conducted with informants.

This Section briefly introduces the notion of collocation as it was developed in two complementary traditions (2.3.1): on the one hand, the British (or neo-Firthian) school of linguistics (Stubbs 1996) – epitomized by the seminal work of John Sinclair – which focuses on collocation as frequent lexical co-occurrence; on the other, the phraseological tradition associated with the lexicographic work of, among others, Mel'čuk (1998) and Cowie (1988), which sees collocation as restricted lexical co-occurrence. The bulk of the chapter is devoted to a discussion of statistical methods for collocation extraction and the lesser known studies that combine corpus-based and psycholinguistic approaches to collocations, which provide the immediate background to the present work.

2.3.1 Two views of collocation: Frequency vs. phraseology

Nesselhauf (2005) makes a distinction between phraseology-oriented and frequency-oriented approaches to the study of collocations which, while not absolute, can help to clarify relevant theoretical and methodological distinctions. Phraseological approaches typically make use of intuition and qualitative observations, and focus their attention specifically on the establishment of criteria for distinguishing collocations from other lexical co-occurrence phenomena and for classifying them into theoretically motivated subsets. Frequency approaches, on the other hand, try to limit the role of intuition in the search for collocations and instead rely on statistical methods for identifying collocations (i.e. frequently used word combinations, regardless of their nature) in corpora of authentic texts. While the distinction is admittedly not always clear-cut – a few studies have attempted to classify collocations extracted from corpora using phraseological criteria, e.g. Nesselhauf (2005), Bartsch (2004) – most researchers do prioritize one or the other view. Given the focus of this study on lexical association measures, the frequency approach is the main frame of reference. Yet the phraseological approach provides relevant insights in terms of parameters of collocations that are worth briefly surveying.

2.3.1.1 The frequency approach

The origins of the frequency approach to collocation are closely associated with the work of Firth and his followers. Firth (1968a:106-107) famously defined the study of collocation as "the study of key-words, pivotal words, leading words, by presenting them in the company they usually keep - that is to say, an element of their meaning is indicated when their habitual word accompaniments are shown". The assumption here is that (part of a) word meaning is established on the syntag-

matic axis, on the basis of relationships existing with co-occurring words. Thus for instance, it would be almost impossible to define the adjective "husky" without a reference to the noun "voice", with which it typically collocates. Scholars following a frequency approach to the study of collocation assume that collocations are not readily available to a speaker's declarative language competence, and therefore are hard to investigate simply tapping into one's own intuition (Xiao and McEnery 2006:103). For this reason methods are needed for extracting collocations from corpora. Early work by Sinclair (1966) proposed that significance of collocation between a node and its collocates be calculated by comparing the actual number of times they co-occur in a given span with their expected probability of co-occurrence (given by the frequency of the node, multiplied by the frequency of the collocate, multiplied by the length of the collocation span in words, divided by the number of words in the text/corpus). Seminal work by Jones and Sinclair (1996) also focused on methodological issues of primary importance for empirical studies of collocations. All arbitrary decisions made in defining what counts as a collocation candidate are discussed, e.g. criteria for the selection of nodes and collocates, length of collocation spans (i.e., the maximum number of intervening words allowed between a node and a collocate), minimum joint frequency of node and collocate, and significance threshold level. As we shall see in the next Section, these (and other) parameters are fundamental for defining collocations within a frequency-oriented study.

Sinclair's work on collocation has mainly focused on specific nodes and their collocates (Sinclair 1991, 1996, 1998). This method consists in identifying one or more node words, and searching for their collocates in a given span around them. This is also the method adopted by Stubbs (2001), who uses it to set up models of extended lexical units around "interesting" words and lemmas. For each word under analysis, a model includes its typical lexical collocates, the grammatical classes with which it tends to colligate, its semantic preference and semantic prosody, and information about distribution and position in texts (Stubbs 2001:87-88).

The keyword method is not the only possible way of searching for collocations in texts. Corpora can also be searched for patterns, i.e. "linear sequence[s] of uninterrupted word-forms [...] which occur more than once in a text or corpus" (Stubbs 2002:230). This is the approach followed by, e.g., Biber and colleagues (Biber et al. 1999) in their work on lexical bundles. Other studies have applied it to units other than words, such as parts of speech (POS). For instance Johansson (1993) retrieves all Adverb-Adjective sequences from a POS-tagged version of the LOB corpus (Johansson et al. 1986) in order to study patterns of adverbial premodification of adjectives (without limiting the search to any specific adverbs/adjectives).

Both methods have (dis)advantages, the keyword method being more subject

to arbitrary choices of the researcher, and the pattern method not allowing modification or expansion of the observed collocations. Whether one or the other is adopted depends on one's own objectives and constraints. I shall discuss this issue further in the next Section, in which the parameters used in the literature for collocation extraction are presented.

2.3.1.2 The phraseological approach

A typical phraseologically-oriented definition of collocations is provided by Howarth (1996:37), who defines them as "fully institutionalised phrases, memorized as wholes and used as conventional form-meaning pairings". Moving from definitions to the actual identification and classification of (authentic) phrases is not straightforward, however, since "institutionalisation" is recognized to be "an intuitive measure" (Howarth 1996:90). A set of parameters is required for classifying collocations and for distinguishing them from other word combinations occurring at the lexical level (i.e., free word combinations and idioms, occupying opposite poles of a cline whose midpoint is occupied by collocations). Howarth suggests that commutability of elements and (non-)literalness can be used to distinguish collocations from free word combinations, while motivation distinguishes collocations from idioms. Similarly, Cowie (1988:131) sees collocation as occupying a middle ground between free, unrestricted, casual word combinations, and idioms, i.e. "combinations whose constant re-use in a fixed form has led to a radical change of meaning". Other researchers mention unpredictability as the central criterion for collocativeness. This is the case with Benson (1985:65), who claims that only "unpredictable combinations" should be mentioned in monolingual collocation dictionaries, and Hausmann (1997:287), for whom collocations differ from idioms because collocations are transparent but unpredictable, while idioms are both opaque and unpredictable.

Focusing on the relationship between members of a collocation, a distinction has been proposed between node and collocator (e.g. Hausmann and Blumenthal (2006)). This is based on the observation that a collocation is always oriented: as suggested by Hausmann and Blumenthal (2006:4), one does not search for the base bachelor starting from the collocator confirmed, rather the contrary. Tutin and Grossmann (2002) attempt to provide a typology of collocations (along the lines of Howarth (1996)) which also incorporates this distinction between node and collocator. In their framework, the salient features would be arbitrariness and unpredictability of the collocator, as well as semantic opaqueness of the whole collocation. The proposed collocation types would be opaque collocations (arbitrary, unpredictable and semantically opaque), transparent collocations (arbitrary and unpredictable, but semantically transparent) and regular collocations (following standard semantic association rules). Mel'čuk (1998:32) provides a

more formalised attempt at describing and classifying collocations based on the notion of lexical functions, i.e. meanings that are expressed lexically in different ways depending on the lexical units to which they refer. For instance, the lexical function **Magn**, meaning "intense(ly), very" is expressed as "strongly" in the context of "condemn", as "pie" in the context of "easy" and as "close" in the context of "shave". By analogy with the more idiosyncratic cases, even expressions like strongly condemn, which are transparent and predictable, are treated as phrasemes (set phrases) in Mel'čuk's framework. In this case therefore (restricted) commutability or (un)predictability are not criteria for collocativeness.

Despite the undeniable interest of these approaches and the value of the insights they provide, it is hard to see how criteria such as predictability, commutability, motivation and transparency could be operationalized in practice (i.e., if one were to apply them to an empirical classification task based on authentic language data). This is in fact not a priority in most studies adopting a phraseological approach, possibly assuming, with Hausmann (1999:127), that all collocations are known to the lexicographer, who merely has to activate her/his dormant competence. This is clearly a major difference with respect to the frequency approach, in which intuitions about lexical syntagmatic relations are viewed as inadequate, and in need of corpus evidence to back them up (Sinclair 1991:4).

2.3.2 Core parameters of collocation

Given the wide range of insights offered in different linguistic traditions, both theoretical and applied, circumscribing the notion of *collocation* and charting the different terms adopted in the literature referring to the underlying concept is a challenging task. On the one hand, the term "collocation" encompasses different phenomena, defined with varying degrees of specification, occurring at the lexical level; on the other hand, lexical phenomena with similar, if not identical, features may be referred to with different names. Examples of these conceptual/terminological divergences and intersections abound: Evert (2005:17), e.g., uses the term "collocation" in an admittedly loose sense ("a generic term whose specific meaning can be narrowed down according to the requirements of a particular research question or application"), while Nesselhauf (2005:Ch.. 2) establishes strict syntactic and semantic criteria for a sequence of words to count as a collocation. Inversely, considerable overlap can be observed between certain category names, as remarked, e.g. by Cowie (1998a:10) with regard to Moon's (1998b) "anomalous collocations" and the term "restricted collocations" adopted in other studies.

This lack of systematicity is widely acknowledged in the field of phraseology – Granger and Paquot (2008:29) speak of the "fuzzy borders of phraseology" – and several authors have identified diverging sets of quantitative and qualitative

criteria according to which collocations are generally defined in the literature (e.g. Seretan 2008; Siepmann 2005). These criteria include frequency-related as well as phraseologically-oriented (i.e. semantic and syntactic) considerations, and provide a level of abstraction by singling out "core" factors, or parameters, of collocativity proposed across specific (sub-)categorizations. Not all the criteria are relevant to all studies: depending on, e.g. whether the approach is more "frequency-" or "phraseology-oriented" (cf. Section 2.3.1), some of them are disregarded. However, they provide a useful practical framework within which to present the variety of approaches to collocation, and a way to systematically "mark the borders" of the phenomenon.

The Sections that follow (2.3.2.1-2.3.2.6) draw in particular on the models proposed by Bartsch (2004:58 ff.) and Gries (2008): most of the parameters described here are derived from these studies, with a few adjustments in terms of the categories' granularity and the order in which they are presented. Parameters include the nature of the linguistics elements involved in collocations, their length and frequency, the "textual" relations among members, and finally lexical, semantic and syntactic properties. While a degree of overlap exists with the previous background Section (2.3.1, here each criterion is presented in turn, together with a brief discussion of ways in which it has been implemented in practice. Presentation proceeds from the more complex and/or central parameters to those that appear to be either more straightforward or more marginal for the approach taken in this thesis.

2.3.2.1 Linguistic elements involved

Three main dichotomies are relevant to the description of this parameter, depending on whether collocations are defined as co-occurrences of *a)* word forms vs. different inflectional variations of the same word form (i.e. lemmas); *b)* lexical, open-class items and/or grammatical, closed-class items; *c)* "higher order", more abstract structures, e.g. the (co-)occurrence of a verb and a particular grammatical construction.

As regards point a), the view that collocation is best described as a combinatorial phenomenon applying to word forms is often justified on the grounds that different inflectional forms of a lexeme may be more or less frequent in actual usage and even have different collocates (e.g Bartsch 2004; Sinclair 1991), in extreme cases giving rise to "different meanings" according to Stubbs (2009:120). On the contrary, when co-occurrence of lemmas is taken into account, this is either because the distinction, e.g. between different tenses of a verb, is not considered relevant, or because the use of lemmas represents an opportunistic strategy to increase the amount of evidence and thus counteract the well-known data sparseness problem that is encountered in studies dealing with phraseology, and infrequent

items in general (Kilgarriff and Tugwell 2002; Seretan 2008:58).

The second dichotomy opposes views according to which only lexical words (nouns, verbs, adjectives and adverbs) can form a collocation, and others which also take into account grammatical, closed-class items (e.g. articles, prepositions, etc.). Studies adopting the first view include Hausmann and Blumenthal (2006) and Mel'čuk (1998), who are especially interested in classifying collocations according to their semantic properties (mainly based on the analyst's introspective judgment), but also more empirically-oriented, corpus-based approaches, such as those in Evert (2005) and Bernardini (2007). The second view has led to the development of notions such as those of *grammatical collocation* (cf. Benson 1989; Carter 1998:58-61), and *collocational frameworks* (Renouf and Sinclair 1991), where the nature of the elements involved in a sequence is not specified *a priori*.

While the first two dichotomies set apart different notions of collocations and/or descriptions of their subcategories (e.g. lexical vs. grammatical collocations), the third seems to mainly discriminate between collocations and related, though distinct, lexical phenomena. When authors deal with the interface between lexis and abstract linguistic structures, they usually prefer to avoid the term "collocation". Phenomena such as "colligations" (Hunston 2001; Sinclair 1991) and the more recently proposed "collostructions" (Stefanowitsch and Gries 2003) seem to lie outside the domain of collocation per se, which usually involves a relationship between lexically specified items.

2.3.2.2 Length of the sequence

A high degree of variation can be observed in the literature concerning how many words constitute a collocation. The widely known definition by Sinclair (1991:17) states that a collocation is a "cooccurrence of two or more words within a short space of each other" in texts. Several authors do not set a specific length *a priori*. This is especially true of qualitative studies (cf. 2.3.1), e.g. Moon (1998b:20) defines "anomalous collocations" simply as "strings", without further specification.

When sequences of words are extracted (semi-)automatically from corpora, length is instead usually defined explicitly: e.g. in their evaluation of statistics-based extraction techniques, Evert and Krenn (2001) and Pearce (2002) restrict the number of words in a sequence to two. Of course, exceptions exist to this tendency, and some NLP-oriented definitions remain vague as to this parameter, e.g. McKeown and Radev (2000:507), who define collocations as "group[s] of words", or Sag et al. (2002:197), who do not mention length at all ("we reserve the term collocation to refer to any statistically significant cooccurrence").

Within corpus-based investigations, sequences longer than two words but with a specified length have been defined with different names, including *lexical bundles* ("Lexical bundles are recurrent expressions [...] of three or more words"

Biber et al. (1999)) and *n-grams* (e.g. Danielsson 2003). Several computational techniques have also been implemented to establish in a bottom-up, data-driven way the "optimal" length of any given sequence to count as a collocation-like construct. Examples can be found in Smadja (1993), Mason (1999), Stubbs (2002) – who calls such sequences *chains* – and Daudaravičius and Marcinkevičiene (2004).

2.3.2.3 Frequency of co-occurrence

In a number of definitions, this criterion lies at the very heart of the concept of collocation. Broadly speaking, the fact that certain word sequences appear repeatedly in texts is taken as an indication of their salience in language competence (cf. Sections 2.3.1 and 2.3.3).

A broad distinction can be made between studies which rely on "raw" frequency of co-occurrence as a criterion of collocativity, and studies in which collocations are defined as sequences of words occurring more frequently *than would be expected*. The former usually include a qualitative evaluation step in which manual selection is carried out to distinguish collocations from casual combinations of frequent words (Moon 1998a; Nesselhauf 2005). The latter try to do without this analytical step by making use of a statistic that compares the individual frequencies of the words considered and their joint frequency. This is a point we will return to in Section 2.3.3.

2.3.2.4 Permissible distance

This parameter refers to the horizon in which collocations are searched for. There are two main ways of applying this parameter. One, defined by Evert (2008:11-15) as "surface co-occurrence", refers to "the permissible distance between the elements involved" (Gries 2008:4), i.e. whether lexical items are adjacent, or appear within a short distance of each other in texts. This distance, measured in terms of tokens intervening between the lexical items of interest, is sometimes referred to as "collocational span". Sinclair (1991:175) proposes a distance of 4 words to the left and right of the node as an optimal span for collocation studies.

The second approach to this parameter has been defined as "textual co-occurrence" (Evert 2008:11-15): collocations consist of lexical items that appear within a pre-defined textual unit, e.g. a sentence, a paragraph, or a whole text. While collocating words are customarily searched for within short textual distances, as Bernardini (2007:33) remarks, related lexical phenomena can span across whole texts, as in the case of Hoey's (2005) *lexical priming*, or Halliday and Hasan's (1976) *lexical chains*. The third distinction mentioned by Evert (2008:11-15), co-occurrence within a pre-defined syntactic pattern, does not refer specifically to distance between lexical items, but rather to structural aspects,

and is therefore dealt with in a separate category (Section 2.3.2.7).

2.3.2.5 Lexical combinatory properties

This criterion is especially relevant to phraseological approaches: for this reason, and since the focus of this thesis is primarily on statistical approaches to collocation (cf. 1), it will only be touched upon briefly here.

Bartsch (2004:60-63) argues that two central notions should be considered, with reference to this parameter, i.e.: *a)* whether lexical restrictions on the selection of collocates are considered as a defining criterion for assigning collocation status and *b)* whether lexical co-selection is seen as a *directional* process.

Criterion a) is taken into account, e.g. in Howarth (1996), Carter (1998) – who speaks of "selectional restrictions" –, and Benson (1989), who calls this property "arbitrariness".

As for criterion b), if directionality is not considered as a property of collocations, then words are seen as mutually "attracting" each other (what Firth (1968b:181) calls "mutual expectancy"), a view shared, e.g. by Danielsson (2003). On the contrary, a number of studies take into account directionality and distinguish between a "semantically independent" basis and a dependent collocate, e.g. Hausmann and Blumenthal (2006) and their notion of "node" and "collocator" (2.3.1.2), Kjellmer's (1991) "right / left predictive" phrases, and Sinclair's (1991) "upward/downward" collocation.

2.3.2.6 Semantic unity and transparency

Like the criterion of "lexical combination" described in Section 2.3.2.5, semantic unity and transparency are mainly relevant for phraseological approaches, and it is beyond the scope of this study to summarize the different positions adopted in the literature (2.3.1.2).

Semantic unity and transparency are mainly used as criteria to set apart "restricted collocations" from free combinations and more "idiom-like" units. The most relevant notion with regard to this feature is that of "(non-)compositionality", which is thoroughly discussed, e.g. in Svensson (2008) and Wulff (2008:Ch. 2).

2.3.2.7 Syntactic structure

The last criterion concerns whether collocations are expected to occur within pre-defined syntactic patterns or not.

Few authors include in their definitions of collocations explicit, theoretically-motivated syntactic considerations. Among these, Cowie (1978:132) defines collocation as a "co-occurrence of two or more lexical items as realizations of structural elements within a given syntactic pattern", and Hausmann (1989:1010) states that

"On appellera collocation la combinaison caractèristique de deux mots dans une des structures suivantes" and continues by listing 7 syntactic structures.

In a majority of cases, specification of a syntactic pattern is motivated by practical concerns, e.g. to delimit the field of investigation for a proof of concept study (Bartsch 2004), or as a way of improving the performance of statistical measures (e.g. Evert (2005) and Seretan (2008)).

2.3.3 Collocation and statistical methods

The role of frequency of co-occurrence in defining collocations has been touched upon several times in this Section, first in the brief discussion of the "frequency approach" initiated by the British School of linguistics (2.3.1.1), and subsequently in the description of the core parameters that have been proposed in the literature to circumscribe the phenomenon (2.3.2.3). It was pointed out that several definitions drawing on the frequency approach have made reference to the related, statistical notion of *higher-than-chance* frequency as a defining criterion for collocation status. In what follows, emphasis will be placed on studies in corpus/computational linguistics that have proposed ways to operationalize this criterion for searching collocations in corpora: this allows us to introduce the framework within which collocation is defined in this thesis.

The idea that collocation may be seen as a probabilistic phenomenon is not new: in Section 2.3.1.1, Sinclair's work on "significant collocations" in the 1960's was mentioned (Sinclair 1966). Along similar lines, Halliday (1961) proposed the following definition:

> Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at n removes (a distance of n lexical items) from an item x, the items a, b, c... Any given item thus enters into a range of collocation, the items with which it is collocated being ranged from more to less probable. (Halliday 1961:261)

Since those early years, probability of word co-occurrence has been implemented in a number of statistical measures for the extraction of collocations from corpora – the first notable example being the work of Berry-Rogghe (1973), who implemented Halliday's definition in a computer algorithm for "[compiling] a list of [...] syntagmatic items" (Berry-Rogghe 1973:103).

McKeown and Radev (2000) remark that the development of increasingly refined statistical tools was driven by applied concerns: e.g. in the work of Choueka (1988), Church and Hanks (1990) and Smadja (1993), lexicography was mentioned as the area in which these measures had their most immediate application. By

way of example, Church and Hanks (1990:27-27) suggested that the statistic they proposed (i.e. Mutual Information) could help lexicographers to "speed up the labor-intensive task of categorizing the concordance lines", by picking out "all the significant patterns [...] and rank[ing] them in order of importance."

These statistical tools – which, following Evert and Krenn (2001) and Pecina (2010), will be referred to as "lexical association measures" in this thesis (henceforth AMs) – have much wider application in contemporary corpus linguistics (see, e.g. McEnery et al. (2006:111-119) and Baker (2006:Ch.5)), as well as in the related fields of computational linguistics and natural language processing (see Sag et al. (2002) and McKeown and Radev (2000)). The term "statistical approach" to collocation has been coined by Partington (1998:15) to refer to the use of AMs for collocation extraction, witnessing to the crucial role these measures have come to play in today's corpus-based studies.

A central concern of this thesis is to assess empirically one of the main assumptions on which the use of AMs rests, i.e. that higher-than chance frequency *in corpora* reflects salience *in the minds* of speakers. Sinclair's (1991:110) formulation of the idiom principle seemed to have psycholinguistic/competence-related implications:

> a language user has available to him or her a large number of semipreconstructed phrases that constitute single choices, even though they might appear to be analysable into segments.

Although lying at the centre of corpus-based studies of collocation and related lexical phenomena (cf. Section 2.3), the hypothesis that lexical regularities observed in a corpus actually *are* "semi-preconstructed phrases" that "a language user has available to him or her", has seldom been be put to test.

By way of example, Bartsch (2004:89) states that:

> It is *postulated* in this study that, whenever co-occurrences of linguistic items are observed with a probability that is statistically higher than chance, the observed co-occurrence is [...] a potential collocation [...]. This is based on the *assumption* that frequently recurrent phenomena must be taken as a reflection of a cognitive reality underlying the organisation and processing of language in the human mind. *(emphasis added)*

Hence, in a number of studies the evaluation of "collocation status" of word sequences extracted from corpora is either carried out on the basis of a researcher's intuitive assessment – an approach that is typical of phraseological approaches (cf. Section 2.3.1.2) – or by means of external evidence, such as, e.g. data sets annotated by a limited number of human informants (as in Evert and Krenn

(2001)), or pre-compiled electronic resources, such as machine-readable dictionaries or terminological resources (cf. Evert 2005:137-140). External evidence is most often used to obtain a "gold standard" against which the extracted collocation candidates can be defined as "true" or "false" collocations (Pearce 2002).

According to Pecina (2010), as many as 57 different AMs have been proposed in the literature, and several articles have been devoted to exploring their statistical properties and their performance in extracting "true collocations" (including Pecina (2010), but cf. also Evert and Krenn (2001), Krenn and Evert (2001), Wiechmann (2008)). Since a *large* number of *human informants* was crucial for the evaluation procedure adopted in the present thesis, only a limited number of AMs could be taken into account, so that the collocation candidates included in the evaluation were not too numerous (cf. Section 3.4.2). These AMs are presented in what follows.

2.3.3.1 The AMs considered in this study

It was decided to focus on four AMs: bare frequency of co-occurrence, which represents the simplest of the four measures, though it has been proved to display similar levels of performance compared to other, more sophisticated statistics; two among the most frequently used AMs in corpus/computational linguistics, and namely Log-likelihood and Mutual Information (Evert (2008:1240), e.g., refers to them as "de-facto standards"); and finally Lexical Gravity, which represents an example of a "linguistically informed" measure, which, despite its potential theoretical and applied interest, has received little attention in the literature. These are briefly described in what follows.

2.3.3.1.1 Frequency of co-occurrence Frequency of co-occurrence (henceforth FQ) is defined as the number of times a collocation candidate is found in a corpus. While, strictly speaking, FQ is not a statistical measure, it was included among the target AMs since it has been demonstrated, when combined with a part-of-speech filter (as is the case in the present work, cf. Section 3.4.1), to outperform more sophisticated measures in a variety of collocation extraction tasks (Krenn and Evert 2001).

2.3.3.1.2 Mutual Information Mutual Information (henceforth MI) was introduced by Church and Hanks (1990), who defined it as follows:

$$MI(x, y) = log_2 \frac{p(xy)}{p(x)p(y)}$$

Here, *p(xy)* is the probability that the two words x and y co-occur, and *p(x)* and *p(y)* represent the individual probabilities that the two words occur

separately. MI is known to give prominence to rare but salient word combinations (Evert 2008): as such it displays the opposite trend compared to FQ.

2.3.3.1.3 Log-likelihood Log-likelihood (henceforth LL) was introduced by Dunning (1993). Since the implementation of LL adopted in this study was derived from Evert (2005:83), the formula below is taken from his work:

$$LL = 2 \sum_{xy} O_{xy} log \frac{O_{xy}}{E_{x,y}}$$

Here, *O_xy* represents the observed frequency of the words x and y and *E_xy* their expected frequency. LL is arguably the most widely used AM in computational linguistics, due to its property of extracting both high- and low- frequency candidates: as such, it occupies a middle ground between FQ and MI (cf. Baker 2006:112).

2.3.3.1.4 Lexical gravity Finally, Lexical Gravity (henceforth LEXG), was introduced by Daudaravičius and Marcinkevičiene (2004), who provided the following formula:

$$LEXG(word_{1|}, word_y) = log(\frac{fq(word_1, word_y) * typesafterword_1}{fqword_1} + \frac{fq(word_1, word_y) * typesbeforeword_y}{fqword_y})$$

Unlike most well-established AMs, LEXG does not only take into account the joint frequency of word x and y and their two overall token frequencies. It also incorporates information on frequency of *type* co-occurrence, i.e. the number of different *types* that co-occur with x in the position of y (and vice versa). This has been suggested to be a relevant criterion, e.g. within phraseological approaches, where restricted collocations are usually defined in terms of the number of different words (i.e. types) a node co-occurs with (cf. Section 2.3.2.5). Despite its potential theoretical interest, this measure is still underexplored in the corpus linguistics literature. Exceptions are Gries (2010a), Gries and Mukherjee (2010), and Ferraresi and Gries (2011).

2.3.4 Collocation and experimental methods

The previous Section introduced the main framework which the present study draws upon to define collocation, i.e. what Partington (1998:15) calls the "statistical approach" to collocation definition and identification. It was argued that the focus on language performance "as a product" (Leech 1992:108) which characterizes this approach to collocation – and corpus-based approaches to language description in general – has often led to overlooking its "process" (or psychological) counterpart, which instead lies at the very centre of the notion, most notably in Sinclair's formulation of the idiom principle (1991; cf. Section 2.3.3).

Gilquin and Gries (2009) have argued that the investigation of the mental processes underlying the production and comprehension of lexical phenomena – as well as other aspects of language, such as syntax and semantics – still seems to be the realm of a different discipline, namely psycholinguistics. The psycholinguistic perspective on "formulaic sequences" – to adopt the coverall term suggested by Wray (2002) – is interested in such questions as *how* they are stored in the minds of speakers (e.g. as a set of freely combinable components vs. single, "prefabricated" units), and *whether* their status as "conventionalized" units makes them "more easily retrieved and processed" than if "the same word sequences were generated through the use of syntax and vocabulary" (Schmitt et al. (2004:128); see also Pawley and Syder (1983), who were among the first authors to put forward this hypothesis, and Wray (2002:Ch.3) for an overview).

Gilquin and Gries (2009:16) remark that despite the relevance of psycholinguistic insights and methods to corpus linguistics, "papers with a corpus-linguistic perspective that combine corpus data with experimental methods [are] rare". The opposite, however, does not hold, and much recent work in psycholinguistics has extensively exploited corpora and corpus-derived data (e.g. lexical databases) as a source of experimental evidence.

Against this background, the present Section aims to provide an overview of the ways in which the methodologies typical of either discipline have been fruitfully combined in lexical research. The range of phenomena investigated is rather wide, and includes two-word sequences as well as larger units (e.g. clusters made of three or more words), "collocations" as defined in the frequency/statistical approach as well as word sequences classified according to phraseological criteria (cf. Section 2.3.2). For ease of presentation, studies will be grouped based on the experimental method they adopt, i.e. word association tasks (2.3.4.1), acceptability judgments questionnaires (2.3.4.2) and lexical decision tasks (2.3.4.3), which were selected among the most widely adopted methodologies in psycholinguistic research on collocational knowledge.[1] This also makes it possible to discuss the applicability of each of them for the purposes of the present thesis.

2.3.4.1 Word association tasks

In its simplest form, the Word Association Task (henceforth WAT) also known as "elicitation test", consists in compiling a list of stimulus words and asking participants to provide "the first word that comes to mind" when faced with the stimulus. As an "off-line" task, i.e. one in which participants responses are not timed, its main aim is that of providing (indirect) evidence as to the storage

[1] Other studies have approached the topic, e.g. through eye-tracking experiments Underwood et al. (2004) or through phonological investigations Pluymaekers et al. (2005) (cf. Shaoul and Westbury (2011) for a review).

of lexical sequences in the mental lexicon, as opposed to their retrieval. As we shall see, among the three methods taken into account in this Section, this is the one that has produced the most controversial results, which has repeatedly led researchers to question its appropriateness for tapping speakers implicit competence. It is nonetheless included in this review as an example of a cued *production* test (unlike the methods described in 2.3.4.2 and 2.3.4.3 below, which exemplify reception/comprehension tests).

The studies presented in Nordquist (2009) and McGee (2009) are examples in which the combination of the WAT methodology and corpus data has provided consistent results across different experimental designs. Nordquist (2009) set up a WAT experiment in which she asked 54 student participants to produce a sentence for each of 12 stimulus words, manually selected from the *Switchboard Corpus* (Godfrey and Holliman 1997). She then compared the responses provided by participants against corpus evidence in two separate analyses: in the first one, she found a significant mismatch between the elicited sentences and *the most frequent corpus collocates* of the stimuli; in the second, more qualitative analysis, she focused on the responses prompted by a single stimulus, i.e. "necessarily", and found that if *structurally complete phrases* extracted from the same corpus are used as benchmark (rather then the most frequent collocates in general), a higher degree of overlap between elicited and corpus data emerged. From this, the author concludes that speakers would seem to "use prefabricated, holistically-stored language in their elicited responses" Nordquist (2009:125).

In a similar experiment, McGee (2009) presented 20 University lecturers with 20 stimulus adjectives, and asked them to write down the *noun* that they thought was the *most frequent collocate* of each stimulus. In this case, too, participants' lexical intuitions differed to a major extent when compared to the most frequent noun collocates in the corpus. Similarly to Nordquist, McGee carried out a more qualitative-oriented inspection of the responses: such analysis revealed a higher degree of overlap between corpus and elicited data in cases where the most frequent adjective-noun sequences occurred in the corpus within "bare dyads". This means, e.g., that participants provided "idea" as a collocate of "good", but did not provide "part" as a collocate of "small"; the author argues that this is due to the fact that "good idea" is a complete, self-contained unit (according to corpus evidence), while "small part" typically occurs within the larger sequence "a small part of". It should be noticed that the different nature of the tasks in the two studies might have influenced the kind of responses provided: in the case of Nordquist (2009), participants were asked to produce sentences, which simulates spontaneous language production, while the task in McGee (2009) prompted more explicit meta-reflection on language. Interestingly, however, the two authors draw similar conclusions, i.e. that the mismatch between corpus and elicited data may be explained by the fact that frequency of co-occurrence alone, as measured in a

corpus, does not capture the form in which formulaic sequences are stored in the mental lexicon.

While providing stimulating insights, the two studies may be seen as somewhat problematic: in both cases, conclusions are based on a (small) subset of the complete set of participants' responses, which leaves unexplained why in the majority of cases corpus and elicited data provide diverging evidence. More methodologically-oriented studies have made an attempt to answer this question.

In a small-scale WAT experiment, Fox (1987) found that most of the responses tended to be of a paradigmatic nature: when presented with a stimulus word, participants were more likely to provide synonyms, or semantically associated words (e.g. "hint" prompted responses like "clue", and "feet" prompted "legs"), rather than the kind of syntagmatic, recurrent patterns which represent the majority of corpus collocates Sinclair (1991). In a similar vein, Gilquin (2008) observed that polysemous verbs like "give" and "take" prompted a concrete interpretation of the verbs' meanings (e.g. "I gave him a chocolate"), while in a corpus the most frequent uses were associated with their abstract, delexical meaning (e.g."And she'll look at me and give me this crazy look"); in passing, it can be noticed that these results also support Sinclair's claim (1991:113) according to which, when tapping speaker's intuitions, the "core" meaning of a word would be its most concrete one.

In the study by Fitzpatrick (2007), corpora play a relatively minor role, serving mainly as a source for compiling a stimulus list – which, as remarked by Gilquin and Gries (2009), is common practice in psycholinguistics. Her results, however, are worth mentioning, insofar as they shed light on an often neglected aspect of WAT experiments, namely the influence on overall results of individual variation among respondents. Aiming to question the appropriateness of WAT methods for tapping native speakers' competence, and of using their responses as a "gold standard" to test non-native speakers' lexical knowledge (as is the case, e.g., in Schmitt (1998) and Granger (1998)), the author demonstrates that, even among native speakers, a high degree of inter-subject variation is found in terms of the number of paradigmatic/syntagmatic responses they supply:[1] while some subjects tend to consistently provide syntagmatic responses, other are more likely to produce paradigmatic ones. Besides challenging the view according to which native speakers' intuitions provide a methodologically sound, homogeneous benchmark against which non-native speakers' responses can be evaluated, the insights afforded by the author on individual variation also call for attention when elicited and corpus data are compared.

[1] Actually, Fitzpatrick (2007) proposes a much more complex categorization scheme of WAT responses. Here, her categorization is presented in oversimplified terms just for the sake of clarity.

Similarly to Fitzpatrick, Mollin (2009) does not specifically aim to tap the mental representations of lexical syntagmatic relations in the minds of speakers. Rather, the author is interested in the relation between word associations (both paradigmatic and syntagmatic) and corpus data. In particular, she compares co-occurrence data from the BNC and a dataset derived from of a large-scale WAT experiment, the *Edinburgh Association Thesaurus*.[1] In her comparison, she considers two main variables that might explain the divergence between speakers' intuition and textual data, i.e. *a)* the word class of the stimulus, and *b)* the lexical association measure that is used to extract corpus collocates. As regards the first variable, Mollin finds that the word class of stimuli tends to have a major influence on the word class of responses (e.g. nouns usually prompt other nouns in the responses), and that the patterns observed in stimulus-response pairings do not always match those found in corpora (in the BNC, e.g. the most frequent word class of collocates co-occurring with nouns is either a preposition or an article). Concerning variable *b)*, the author suggests that some statistical measures (i.e. MI and z-score) are better able to extract collocation candidates that are also provided as responses in a WAT. As was also noticed by Nordquist (2009) and McGee (2009), frequency (as well as other measures that are highly correlated with frequency, e.g. Log-Likelihood) proves instead a bad predictor of word associations, mainly due to the fact that it tends to give prominence to function words (which are rarely, if ever, provided as responses in WATs; cf. Clark (1970)). This point will be further elaborated upon at the end of Section 2.3.4.2.

2.3.4.2 Acceptability judgement questionnaires

Compared to WAT experiments, acceptability judgement questionnaires (henceforth AJCs) seem to be a better-established methodology for tapping speakers' intuition and collocational knowledge (Murphy 2007:13,62-63). AJCs are off-line tasks in which participants are presented with a list of stimuli (e.g. word combinations), and are asked to "evaluate" them according to a specific criterion, e.g. their perceived degree of acceptability (Granger 1998), or the strength of association between words (Lapata et al. 1999). Judgements are either required in the form of a numerical value (e.g. a scale ranging from negative to positive values), or making reference to descriptive labels (e.g. "totally unacceptable", "no opinion", "totally acceptable"). In what follows I will distinguish between two different ways in which corpora complemented research designs involving AJCs, i.e.: *a)* corpora served as a benchmark to test a psycholinguistic hypothesis; *b)* corpus and experimental data were "on an equal footing" Gilquin and Gries (2009:11), i.e. the two types of evidence were used to validate each other.

[1] http://www.eat.rl.ac.uk [Last consulted 29.11.11]

The work of Siyanova and Schmitt (2008) exemplifies the use of corpora as a benchmark. The authors present an AJQ experiment comparing English native speakers' and learners' intuition on adjective-noun collocations, through which they test the hypothesis that the two groups differ significantly in terms of the accuracy of their collocational judgements (a further experiment was carried out involving a lexical decision task; this is discussed separately in 2.3.4.3 below). The gold standard against which responses were evaluated was represented by a list of "unambiguously appropriate" collocations (Siyanova and Schmitt 2008:440), manually selected among adjective-noun sequences in the BNC which displayed a medium or high frequency of co-occurrence, a MI value higher than 3, and which were included in two collocation dictionaries; implausible word sequences were added to the AJC stimulus list as control items. Results indicated that both native and non-native speakers' had reliable intuitions in telling apart plausible vs. implausible collocations, although native speakers tended to provide more "extreme" values than learners, i.e. plausible collocations were scored higher (and implausible collocations were scored lower) by the former group than by the latter; moreover, while native speakers' scores for high-frequency word pairs were significantly higher than those for medium-frequency ones, no significant difference was found for the group of learners. This is interpreted as evidence that "their knowledge was just not as accurate as that of the [native speaker]" (Siyanova and Schmitt 2008:445). Interestingly, these results seem to be consistent with those of Granger (1998), who, in a similar AJQ experiment also observed that learners perceive as "particularly salient" (Granger 1998:152) fewer collocation types than native speakers.

The studies presented in Lapata et al. (1999), and Ellis and Simpson-Vlach (2009) take a more exploratory approach to the relationship between corpus and experimental data. A fundamental difference sets apart the research design adopted in these studies from that of Siyanova and Schmitt (2008): unlike the latter, the former do not assume that corpora can provide "unambiguously appropriate" collocates; rather, they start from the premise that different corpus linguistics metrics of formulaicity may affect the accuracy of processing of formulas in native speakers (Ellis and Simpson-Vlach 2009:61). Given the relative lack of studies exploring the interface between corpus and psycholinguistic data, this seems a fair assumption, and one that might explain why the use of corpus-derived data as gold standards in psycholinguistic studies is relatively uncommon (Gilquin and Gries 2009:14).

Going back to the studies by Lapata et al. (1999), and Ellis and Simpson-Vlach (2009), both of them aim to assess the extent to which "metrics of formulaicity" correlate with human judgements collected through AJQs (as in the case of Siyanova and Schmitt, Ellis and Simpson-Vlach also set up a lexical decision task, which will be discussed in Section 2.3.4.3). The ways in which stimuli are

selected and the metrics taken into account, however, differ: Lapata et al. (1999) focus on two-word combinations in the adjective-noun syntactic pattern; these are extracted from the BNC from three frequency bands (high, low and medium), and undergo a process of (mainly automatic) pruning (e.g. the head noun had to have a frequency of at least 10 occurrences per million word); finally, the remaining pairs are scored according to five measures of lexical association, including bare frequency and log-likelihood. On the other hand, Ellis and Simpson-Vlach (2009) take into account longer word sequences (3- to 5-grams), and two lexical association measures, i.e. frequency and MI; stimuli are then selected by stratified random sampling (cf. also Section 3.4.2) to represent three levels of each of these variables (e.g. high/medium/low frequency pairs, pairs with a high/medium/low MI score). The lists of stimuli thus compiled were then submitted in the form of an AJQ to informants, who were asked to judge the "goodness of fit" of the word combinations (Lapata et al. 1999) or their "teaching worth" (Ellis and Simpson-Vlach 2009). Unfortunately, only frequency of co-occurrence is taken into account by both studies, and hence results can be directly compared only with regard to this feature. Both Lapata et al. (1999) and Ellis and Simpson-Vlach (2009) found frequency of co-occurrence to be significantly correlated with human judgements: however, while the former observed that it was the strongest predictor, the latter found that it was outperformed by MI (and the same trend, as we shall see, was observed in the lexical decision task).

Despite the apparent similarity of the research designs just described, one crucial difference can be noticed between the ways in which stimuli were selected in the two studies, which seems to be related to the different perspectives from which data are approached. The motivation underlying the work by Lapata and her colleagues was an applied one: they extracted collocation candidates following established practices in corpus linguistics, e.g. by applying a part-of-speech filter and setting frequency thresholds (notice that no *manual* selection of stimuli was carried out), thus simulating a practical collocation extraction task. On the other hand, Ellis and Simpson-Vlach took a somewhat less refined approach to collocation extraction: i.e., these two authors extracted high-frequency n-grams irrespective of syntactic patterns, or n-grams with *low* MI scores. This is done to ensure that examples of all combinations of the variables considered are included in the data set (e.g. pairs with high frequency/low MI; pairs with low frequency/high MI, etc.), following a common procedure in *psycholinguistic* studies (the so-called "cross-tabulation" of variables), which is not as common in corpus linguistics. It should be remarked that the word sequences sampled in this way included, e.g., combinations of function words or grammatical collocations, which are unlikely to be perceived as interesting by human informants (e.g. "is one of the", "the content of"; cf. also Mollin (2009)). One might therefore legitimately wonder whether the fact that frequency was outperformed by MI is a result of

the "properties" of frequency per se, or rather of the sampling method adopted, and whether Ellis and Simpson-Vlach's (2009) results can be said to have direct implications for more corpus linguistics-oriented applications (e.g. to decide on an appropriate measure for collocation extraction).

2.3.4.3 Lexical decision tasks

The last experimental procedure that will be taken into account is the lexical decision task (henceforth LDT). This is an example of an-online task, in which informants are presented on a computer screen with a series of stimuli and are asked to press one of two keys according to whether, e.g. they perceive the stimulus as a plausible word combination or not: their answer and reaction time (RT), i.e. the time elapsed between the onset of the stimulus on the screen and the moment they press a key, are recorded. The accuracy with which respondents recognize experimental items as plausible and, most of all, their reaction times are assumed to be directly related to retrieval strategies in the mental lexicon: in broad terms, the higher the accuracy level, and the shorter the RT, the more it is likely that a word pair is stored in, and retrieved from, memory as a whole unit, which results in faster processing/recognition (Ellis 2002). In what follows, studies combining LDTs and corpus data will be presented: in this case, too, a distinction will be drawn between hypothesis testing approaches to corpus/experimental data, and exploratory ones.

As was mentioned in Section 2.3.4.2, Siyanova and Schmitt (2008) used the same data set of the AJQ experiment in a LDT to test a related aspect of their hypothesis, i.e. that not only do native and non-native speakers display different levels of collocational knowledge when explicitly prompted to evaluate word combinations, but also that the two groups differ in terms of *how quickly* they are able to recognize the same combinations as plausible or implausible. Interestingly, the results obtained in the LDT pointed in the same direction as those of the first experiment: non-native speakers had slower RTs than natives, and no significant difference emerged between the RTs associated with medium- and high-frequency combinations (while such difference was found to be significant for the native speakers' group). According to the authors, this suggests that "not only are NNS judgements of [collocations] less accurate than those of NSs [...] but that the recognition processing necessary to reach those judgements proceeds more slowly for NNSs" (Siyanova and Schmitt 2008:451).

Again, it can be argued that these results are heavily dependent on the specific operationalization of collocation adopted in the study (frequent word sequences with a MI score higher than 3; cf. 2.3.4.2 above): starting from the assumption that corpora can be used to extract "unambiguously appropriate" collocations, the authors disregarded a number of variables that might have influenced their

results. By way of example, Ellis et al. (2008) found that RTs of non-native speakers in a similar LDT tend to be correlated with frequency, rather than with MI, and Wolter and Gyllstad (2011) observed no significant difference between the RTs of native and non-native speakers when collocations in the L2 of the latter have a direct translation equivalent in their L1. If anything, this might suggest that until a clearer picture emerges of the interrelationship between corpus and experimental evidence, the assumption that corpora are valid "gold standards" in psycholinguistic studies might lead to undue generalizations.

For this reason, the exploratory approach adopted by Ellis and Simpson-Vlach (2009, and Ellis et al. (2008); cf. 2.3.4.2 above) would seem to be, at this stage, preferable from a methodological point of view. As in the case of Siyanova and Schmitt (2008), Ellis and Simpson-Vlach (2009) combined in the same study an AJQ and an LDT experiment, finding that the two types of evidence lent support to each other: native speakers' RTs of formulaic sequences were found to be correlated with MI, but not with frequency. This, however, is not hypothesized to be evidence that e.g. pairs with high MI scores *are* collocations, but rather that MI seems to be better able to predict human processing.

2.3.4.4 A combined approach to the evaluation of collocation

Sections 2.3.4.1-2.3.4.3 provided a number of insights into the relationship between corpus and experimental data, that I will briefly summarize here (cf. also Gilquin and Gries (2009)):

- different experimental designs have been be used to tap the relation between a product-oriented view of collocations and a process-oriented one, affording different perspectives on issues such as, e.g. the mental processes underlying production vs. comprehension of collocations, their storage in the mental lexicon vs. their retrieval. Some experimental methods, most notably AJQs and LDTs, have been suggested to provide less controversial results than others (like WATs); moreover, it is not uncommon for the two types of experiment to be combined;

- psycholinguistic studies tend not to make full use of procedures that are common practice in corpus linguistics, such as, e.g. the exploitation of part-of-speech tagging and/or lexical association measures;

- on the contrary, manipulation of experimental variables through careful, manual selection of corpus-derived stimuli is common in psycholinguistic-oriented approaches (see also Gilquin and Gries 2009:8): this makes it possible to minimize the effects on participants' responses of unrelated variables (e.g. word length). At the same time, however, this makes it

harder to evaluate the relevance of results for applied, corpus linguistic-oriented approaches, where no *a priori* selection of data is usually carried out (cf. Leech's (1992:112) "principle of total accountability"). It was suggested, e.g., that even if the selection of stimuli is based on "objective" criteria (e.g. frequency thresholds, MI scores), different sampling strategies, inspired to psycholinguistic vs. corpus linguistic methodological stances, might lead to diverging results, with different implications for the two disciplines;

- finally, the combination of corpus and experimental data can be approached either form a hypothesis testing or from an exploratory or perspective, a distinction which is also closely associated with the different traditions of corpus and psycholinguistics.

With particular reference to the last point, it should be remarked that the study presented in the next Chapters will approach experimental data in an exploratory fashion, and with corpus linguistics-oriented applications in mind. As such, it will not aim to test a specific hypothesis on the mental processes underlying the production or comprehension of collocations. Rather, along the lines of Lapata et al. (1999), it will start from a product-oriented, statistical definition of collocation, and will draw on psycholinguistic methods to try and shed light on the relations between such a definition of collocation and a process-oriented one.

As a conclusion, a passage from Gilquin and Gries (2009) can be quoted:

> Corpus linguists have been developing different quantitative measures of collocational attraction [...]. However, there is comparatively little work that attempts to validate, say, the 20+ collocational measures [...] against findings from corpus-external data and show what, if anything, these measures mean, indicate, or reflect. (Gilquin and Gries 2009:17)

This is the challenge that the present thesis aims to take up.

2.4 Summing up

This chapter has provided an overview of the theoretical background of the present thesis. It has suggested that institutional academic English in general, and degree course descriptions published on the web in particular, make interesting yet still relatively underexplored objects of corpus-based linguistic research. It has also introduced the notion of collocation and discussed several aspects of

special relevance to the present work, in particular statistical methods applied to the identification of collocations in corpora and experimental methods employed to bridge the gap between the product-oriented, performance-based approach to collocation research taken within corpus linguistics, and the process-oriented, competence-based approach typical of psycholinguistics. The next chapter moves on to describe the methodological underpennings of this thesis.

Chapter 3

Corpus and experimental setup

3.1 Overview of the Chapter

This chapter states the research questions addressed by the thesis (3.2), and describes the methodology adopted to answer them. Section 3.3 illustrates the semi-automatic methods used for corpus construction, and describes the resulting corpus of degree course description published on the web by British universities (UniCoDe_UK). Section 3.4 then describes the procedure used to extract the set of collocations subsequently used in the evaluation experiments. This is a crucial step for ensuring that any results obtained on the basis of a (necessarily small) set of observations are unbiased with respect to the AMs being compared, and at the same time that they are representative of the wider data set from which they are extracted. Section 3.5 goes on to present the methodology and results of three separate evaluation experiments, in which the selected collocation candidates were evaluated with respect to *a)* dictionary coverage – taken to represent actual usability for practical purposes – *b)* collocativity judgments provided by expert informants (both native and non-native corpus linguists), and *c)* the implicit knowledge of native speaker informants. The final Section (3.6) introduces the statistical tests that will be used in the next Chapter for the analysis of collected data.

3.2 Research questions

In Section 2.3.3 I argued that most corpus-based studies, and especially those adopting a frequency/statistical approach, *assume* that scores of collocativity reflect the psychological salience of the word sequences under consideration: their higher-than-chance frequency of co-occurrence is taken as (indirect) evidence of their storage as units in the speakers' minds. Furthermore, evaluation methods

often require a binary, "yes/no" classification of sequences into collocations and non-collocations, which is at odds with the view of collocativity as a scalar property (cf. Section 2.3.3). On the other hand, few studies have attempted to assess empirically the degree of overlap between corpus and experimental evidence on collocation (2.3.4): the bulk of these studies have adopted a hypothesis-testing stance, typical of the psycholinguistic approach, which, it was argued (cf. Section 2.3.4.4), has limited explanatory power in "show[ing] what, if anything, [lexical association] measures mean, indicate, or reflect" (Gilquin and Gries 2009:17).

Given the lack of an uncontroversial, widely agreed-upon definition (and evaluation method) of collocation status, in this thesis I concur with the idea of Wray (2002:66) that collocation is "not a single and unified phenomenon", and that "several baselines" are needed to account for what is (or is not) collocational. Focusing on four statistical measures for the extraction of collocations from corpora, i.e. frequency of co-occurrence (FQ), Lexical Gravity (LEXG), Log-Likelihood (LL) and Mutual Information (MI), and viewing psychological salience as a fundamental touchstone against which they should be evaluated, this thesis aims to answer the following three interrelated questions:

1. Do AMs predict experts' intuitions on the salience of a collocation? And if so, does a given AM predict them better than the others? In turn, this also raises the question whether consensus emerges as to what constitutes a salient collocation.

2. Do AMs predict the strength of association of word sequences in the minds of native speakers? And if so, does a given AM predict it better than the others?

3. Finally, to what extent do corpus data, expert judgments, and experimental evidence provide converging evidence as to the phenomenon of collocation?

It should be remarked that the term "collocation" and "collocativity" will be used in a rather loose sense to indicate the degree to which "corpus-external" evidence (e.g. explicit endorsement by experts) testifies to the salience of the word sequences under analysis: as such, it does not refer to a specific notion of collocation as is typical of phraseological approaches (cf. Section 2.3.1), nor does it imply a binary classification of word pairs into collocations and non-collocations (cf. Section 2.3.3).

3.3 Corpus setup

3.3.1 (Semi-)automatic methods of corpus construction: rationale

Given the ever increasing availability of texts in electronic form on the Web, more and more researchers are turning to it as they would to a corpus *shop* (Bernardini et al. 2006). In other words, they query a traditional search engine, such as *Google*, for combinations of search terms and phrases, and download the texts retrieved by the engine, saving them locally and then querying them with a concordancer of their choice. This procedure can be automatized to various degrees, and tools have been developed for this purpose (e.g. the *BootCaT* toolkit used in this thesis (Baroni and Bernardini 2004) and Bill Fletcher's *webascorpus*[1] tool).

Using these tools, a corpus of over one million words can be constructed in a few minutes. Apart from (larger) size, using automated methods has the advantage of making the corpus construction process less subjective, and therefore more easily replicable for other languages, or at a later time. It is sufficient to keep track of the seed words and parameters used, and one can compile a comparable corpus for a different population (e.g., course descriptions in English from universities in other European countries).

But automatization comes at a cost, and the process of retrieving pages matching a given query or set of queries automatically, cleaning them of html code and repetitive text, and saving them locally, if done without human supervision, can result in a corpus that matches the population one was targeting to a very limited degree. While this result can be adequate for some pratical applications (see e.g. Fantinuoli (2006)), it would not be acceptable for the research purposes envisaged in this thesis, which are both methodological and descriptive.

3.3.2 Corpus construction: a step by step account

The corpus construction procedure used in this work therefore combined automatic methods for text retrieval (i.e. the widely used *BootCaT* toolkit (Baroni and Bernardini 2004)) with manual checks, inspection of downloaded pages, and cleaning. This semi-automatic process aimed to strike a balance in terms of the quality/quantity trade-off inherent in all corpus construction activities: a corpus built completely manually is inevitably smaller but cleaner and more representative of the target population, while the opposite holds for automatic corpora (Bernardini and Ferraresi forthcoming).

[1] http://webascorpus.org/ [Last consulted 29.11.11]

The first step of the BootCaT procedure consists in manually identifying relevant "seeds", i.e. words or word combinations that are assumed to be characteristic of the language variety of interest. For research focusing on topic/domain-specific varieties of a language, seeds are usually key terms of that domain (Wong et al. 2011). In the present case, however, the main criterion was different. Rather than sharing a certain topic or domain, they had to belong to a specific *genre*. Therefore an alternative course of action was taken.

The websites of all UK universities with more than 1,000 students were inspected – a total of 147 websites.[1] Based on preliminary inspections, the websites of universities with fewer students often returned very few usable/relevant webpages, for which reason they were excluded. For every remaining website I checked whether course descriptions shared the same "base URL", i.e. a high-level URL that all course descriptions belonged to. To give an example, all undergraduate courses at the University of Manchester share the base URL http://www.manchester.ac.uk/undergraduate/courses/atoz/course/, while the last part varies. Adding "?code=03512" to the base address one accesses the course description of the BSc in International Management, while "?code=00306" finds the BA in French and Spanish (cf. Figure 3.1). If such a URL existed, it could be used as a query to the search engine, that only targeted BA course descriptions. If no such base URL could be found, the whole website/university was discarded. This was an opportunistic decision motivated by the semi-automatic procedure of text retrieval adopted, which, however, was not expected to introduce a bias in corpus composition. Based on random checks, no difference could be gleaned between selected vs. discarded websites in terms of, e.g. the prestige, size, etc. of the respective universities.

A methodological note is in order at this point. The decision to seed searches using URLs has a double advantage. On the one hand, as just discussed, it is likely to reduce the number of incorrectly identified texts that are inevitably included in a fully automatic corpus. On the other, it is a theoretically sound procedure since, as suggested by Elena Tognini Bonelli at a debate held in conjunction with the ICAME 32 Conference,[2] URLs can be considered as external criteria, and thus appropriate parameters on which to base text selection for inclusion in a corpus (Sinclair 2004).

At this point, two strategic decisions had to be made, i.e. *a.* whether to include both undergraduate *and* postgraduate degree course or either of them, and *b.* whether to employ a "stratified sampling" procedure to ensure variety in terms of the disciplinary areas of the various courses. As regards point *a.* I settled to

[1] The list was obtained from Wikipedia http://en.wikipedia.org/wiki/List_of_UK_universities_by_size [Last consulted 12.01.2011].

[2] "Do we still need language corpora?", pre-conference debate organized by Martin Wynne and Ylva Berglund Prytz; Oslo, June 1, 2011.

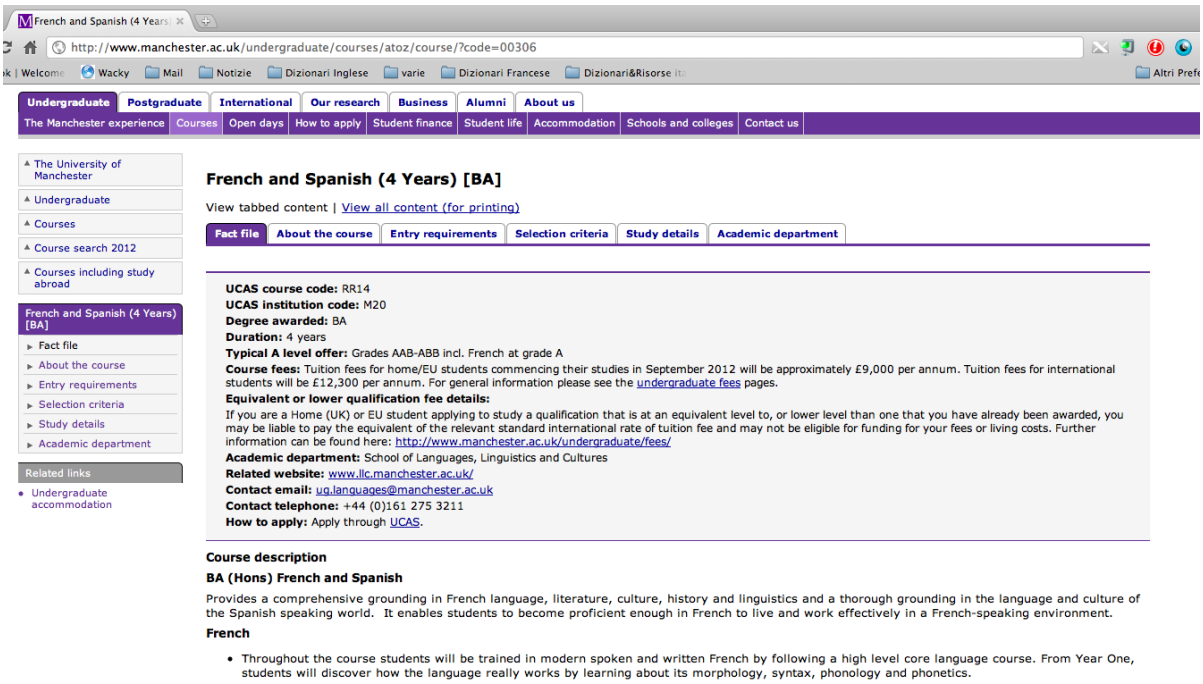


Figure 3.1: A degree course description: BA in French and Spanish at the University of Manchester.

narrow down the selection to undergraduate courses only, which tend to be more homogeneous compared to MA's/MSc's (e.g., the latter include taught and research courses, which are radically different in terms of structure and content) as well as more numerous. URL syntax was once more adopted as a criterion for inclusion/exclusion of the websites: only those for which it was possible to distinguish between undergraduate and postgraduate courses were kept (cf. the example of Manchester university base URL above, that includes the word "undergraduate"). As for the decision on point *b.*, I resolved to adopt no *a priori* criterion for selecting texts, for several reasons. At the theoretical level, every decision as to which disciplinary areas (and sub-areas) to include and in what proportions would have been highly arbitrary. Moreover, based on the methodology adopted for the identification of texts, sampling according to stratified criteria would have implied using topic, i.e. a text internal feature, as a guiding principle (cf. the discussion in Sinclair (2004)).

Once the base URLs had been identified, these were used as arguments of the `site:` operator and submitted to the *Yahoo!* search engine (e.g.: `site:http://www.manchester.ac.uk/undergraduate/courses/atoz/course/`). This ensured that only pages from the subset sharing the given base URL would be

retrieved from the website of the University of Manchester). A maximum of 100 documents per site matching these parameters were downloaded (such was the limit imposed by the search engine). Since the procedure relies on search engines' ranking algorithms, and since I downloaded the first 100 pages, sampling was carried out largely based on the way a particular website is indexed by the search engine itself. It is therefore likely that the pages that ended up in the corpus are skewed towards the more "popular" ones (Gatto 2009:51-52). This is not considered a problem, however, since *a.* such pages have a higher reception status (on the notion of reception status see Burnard (1995)), i.e. they are read by a wider audience, and *b)* the same retrieving procedure is used for all the universities.

A random sample of the URLs returned by *Yahoo!* was then inspected manually for each website. Single webpages or entire websites were discarded at this stage if:

- the search engine yielded no result for the university's base URL;
- pages turned out to be "splash pages", containing only *lists of courses* rather than their descriptions;
- pages were contained little connected text. Admittedly this is an "internal criterion", but including such pages in the corpus would have resulted in no evident advantage either in terms of phraseological items extractable from them, nor in terms of a better understanding of course descriptions as a genre.

In the final phase, the texts were downloaded and post-processed. First, they underwent a process known as "boilerplate-stripping" (Fletcher 2004), i.e. an algorithm was used to remove all those parts which tend to be the same across many pages (e.g. disclaimers, headers, footers, navigation bars, etc.), and which are poor in human-produced connected text (cf. also Baroni et al. (2009)). While relevant for studies of web communication in general, these contents are not strictly speaking part of the genre under consideration, and they are by definition very frequent, since they are repeated *verbatim* on every page. Their inclusion would have yielded more noisy results, with no obvious advantage for the purposes of this study.

Lastly, basic metadata (information on URL and publishing university) were added to the pages, which were then annotated with POS-information and lemmatised using the TreeTagger[1]. As a very last step, the corpus was indexed for

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ [Last consulted: 20.11.11]

Table 3.1: Basic information on the UniCoDe_UK corpus.

number of texts	5,353
number of universities	86
number of tokens	4,837,356
number of types	56,614
type/token ratio	1.17
average text length	903.67
standard deviation	818.917
longest text (tokens)	91,475
shortest text (tokens)	14

corpus consultation with the CorpusWorkBench.[1]

3.3.3 Corpus data

Table 3.1 presents basic data about the UniCoDe_UK corpus.

3.4 Establishment of data set

The following Sections describe in detail the procedure that was adopted to extract and score collocation candidates from the UniCoDe_UK corpus (3.4.1), the sampling method used to select data for the experiments (3.4.2) and concludes by presenting the final data set (3.4.3).

3.4.1 Collocation candidate extraction

In the selection of collocation candidates, a number of decisions had to be made as to the core parameters defining target collocations. Following the classification proposed in Sections 2.3.2.1 to 2.3.2.7, the collocation candidates were defined as follows:

- **Linguistic elements involved:** sequences of word forms, rather than lemmas, are considered. This decision is both theoretically justified – since different word forms might display different collocational patterns (Sinclair 1991) – and methodologically safer, since it eliminates the risk of introducing errors due to the automatic lemmatization process.

[1] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/ [Last consulted: 20.11.11]

- **Length of the sequence:** only two-word sequences are taken into account. This is mainly justified by the fact that standard AMs are usually devised to calculate association scores between two words only. Computation of AMs for longer sequences requires that further methodological decisions be taken,[1] thus adding a level of complexity that should only be adressed once the more standard case (i.e. the binary relation) has been thoroughly understood.

- **Frequency of co-occurrence:** a minimum frequency threshold of at least five occurrences was set. This followed from preliminary investigation of the scored pairs, which showed that MI yielded intuitively implausible results, e.g. it consistently gave prominence to highly idiosyncratic, or "malformed", pairs (e.g. *rigorous btec*, *recentinternational relationsplacements*). This is due to its tendency to highlight low-frequency pairs (cf. Section 2.3.3). The threshold was set to 5 following Church and Hanks (1990).

- **Nature of the co-occurrence:** *all* sequences of two words (in a predefined syntactic pattern, see the last point of this list) were considered as collocation candidates: the approach that is taken here does not imply the selection of keywords and a search for their collocates, since this could introduce personal biases in the final data set. Furthermore, only adjacent pairs are taken into account. While augmenting the collocational span would result in a higher number of collocation candidates and higher joint frequencies (and thus possibly in more accurate estimates of their association strength), it would also increase the probabilities that malformed pairs are introduced in the data set, e.g. words that do not occur in the same phrase, but rather are syntactically related to a third component in the sentence.

- **Lexical combinatory properties / Semantic unity and transparency:** the collocation candidates were not evaluated according to these parameters, which, as argued in Section 2.3.1, largely depend on the evaluator's subjective interpretation.

- **Syntactic structure:** all the pairs extracted belong to the *adjective-noun* pattern. Since the extraction procedure relied on Part-Of-Speech sequences, rather than on syntactic dependency information (which would require dependency parsing of the corpus, as in the approach taken, e.g. by Seretan (2008)), there is no guarantee that the extracted words actually occur in the same phrase. However, reliance on POS sequences is a fairly standard method in collocation extraction, and is adopted, among others, by Evert

[1] E.g. given 3 words a, b and c should one compute association of a with b and b with c or also of a with c? See also Daudaravičius and Marcinkevičiene (2004) on this.

and Krenn (2001), and in the commercial lexicographic tool SketchEngine.[1] To maximise the probabilities that the adjective-noun pairs were actually part of the same noun phrase, the further constraint was imposed that the noun in the pair should not be followed by another noun (so that, e.g. a word sequence like "optional field", always occurring in the corpus within the larger phrase "optional field trip", is filtered out of the collocation candidate list).

The candidate pairs were extracted in the form of a frequency list through the `cwb-scan-corpus` utility (included as part of the CWB corpus manager), for a total of 22,604 word pairs. To reduce noise in the data (e.g. malformed words, characters with wrong encodings), all pairs in which one or both words contained non-alphabetic characters (apart from dashes and apostrophes) were filtered out.[2]. After this filtering stage, the number of word pairs with frequency higher than 5 was reduced to 7,323. These were lowercased and uniqued, and subsequently scored according to the three AMs introduced in Section 2.3.3, i.e. Lexical Gravity (LEXG), Mutual Information (MI) and Log-Likelihood (LL). FQ values were obtained directly from the frequency list; MI and LL scores were computed using a Perl wrapper to Stefan Evert's UCS toolkit;[3] since the toolkit does not include options to compute LEXG values, these were calculated through a Perl implementation of Daudaravičius and Marcinkevičiene's (2004) formula.[4]

3.4.2 Sampling strategy and rationale

The data set for the evaluation of AMs usually consists in a list of the pairs that obtain the highest values according to a given AM, i.e. the *collocation candidates*. This list, also called n-best list (Evert 2005) – where n- stands for any arbitrarily chosen number of collocation candidates (e.g. 100-best list) –, thus contains the n- pairs that an AM selects as being the most "collocational".[5]

The procedure, however, has two main drawbacks. First, as argued by Evert and Krenn (2001), if only the top n- pairs from the scored lists are taken into account the results of the evaluation might not be reliable, i.e. they may not reflect accurately the overall precision of the AM under consideration. According to the

[1] http://www.sketchengine.co.uk/ [Last consulted 29.11.11]

[2] The problem of noise is particularly relevant when corpora are built semi-automatically starting from web data. See Fairon et al. (2007) and Section 3.3.1 for a discussion on this point.

[3] http://www.collocations.de/software.html [Last consulted 29.11.11]

[4] The set of Perl scripts developed specifically for the purposes of this thesis are available from the author on request.

[5] Here I am disregarding the "threshold" and "ranking" approaches described in Evert (2008): this is done both for simplicity's sake, and because n-best lists are the most widely adopted method for collocation extraction and evaluation (Evert and Krenn 2001).

two authors, "results are unstable for the first few percent of the data", since "they are more susceptible to random variation" (Evert and Krenn 2001:41-42). This is best illustrated with a graph: Figure 3.2, taken from Evert and Krenn (2001) shows that all the AMs, including LL and MI, display "random fluctuations" in terms of precision if only few "top" pairs are considered: these are (conjecturally) represented by the portions of the lines enclosed within the black rectangle.

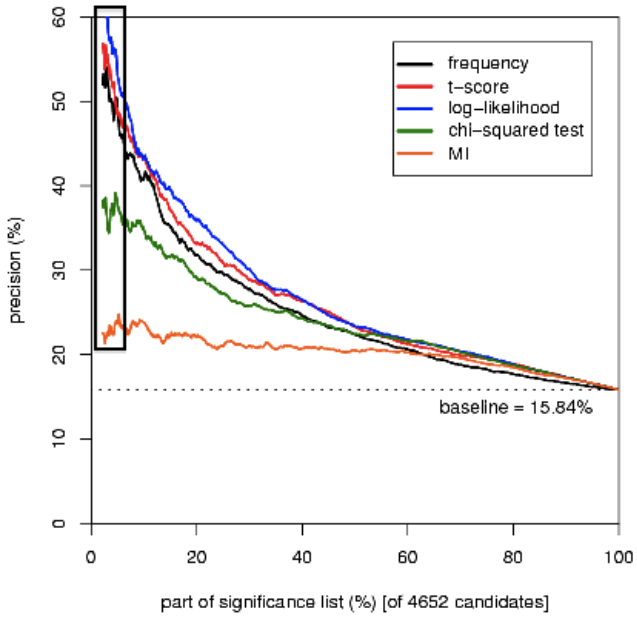


Figure 3.2: Precision curves for different AMs (adapted from Evert and Krenn (2001:42)).

In order to counteract this effect, Evert and Krenn (2001) suggest that a larger proportion of the collocation candidate list, i.e. around 50%, be considered for the evaluation. This, however, would be impractical for the purposes of the present study, which aims at obtaining the greatest possible number of judgements for the collocation candidates, and must therefore limit the data set to be evaluated to a reasonable number of pairs. Since I am interested in comparing four AMs, consistently increasing the number of candidates for each AM even by 10 pairs would result in an increase of 40 units in the data set.

The second pitfall connected with n-best lists is that, if the AM employed tends to extract as collocation candidates only high- *or* low-frequency items, pairs belonging in a different frequency range will be systematically excluded from the

list. This effect is undesirable, since different frequency ranges are known to yield different types of collocation candidates (e.g. pairs that may be considered as "typical" of the domain under consideration vs. rare but potentially salient ones; cf. Bartsch (2004)). It is widely reported in the literature that both MI and LL are characterized by such skewness, giving prominence respectively to pairs with low and high frequency of co-occurrence (cf. Section 2.3.3.1). To the best of my knowledge, no such analysis has been carried out for LEXG.

Thus, prior to deciding on a sampling strategy for the evaluation of the AMs, a preliminary analysis was carried out to investigate the relationship between the AM scores of the adjective-noun pairs from the UniCoDe-UK corpus and their frequency of co-occurrence. Scatterplots were produced to represent graphically the correlation (Figure 3.3). Each dot represents a pair in the complete data set: the x axis represents the logarithmic value of its frequency,[1] and the y axis its score according to the different AMs.

Figures 3.3b and 3.3c illustrate the typicalities of MI and LL just mentioned: as can be observed, the highest MI values concentrate around the left part of the plot, i.e. are skewed toward low frequencies, while the opposite holds for LL. As for LEXG, a distribution similar to that of LL is evidenced. As frequency increases, so do LEXG scores.

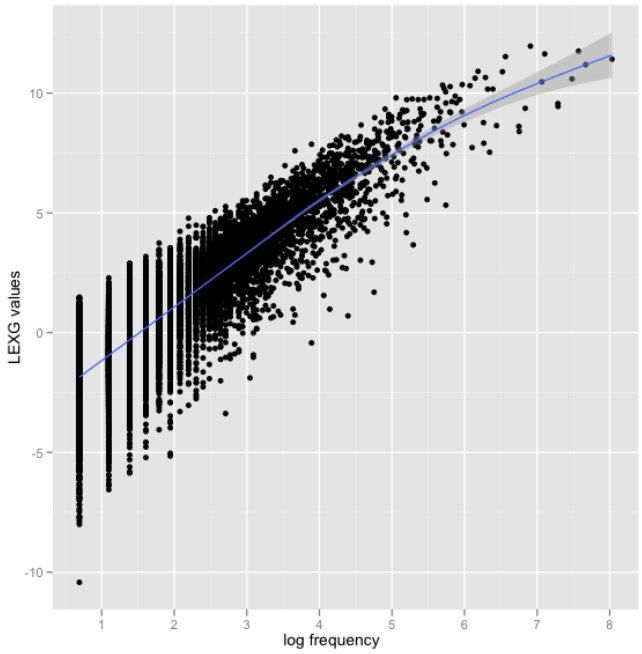
To quantify the correlation between the AMs and frequency, a statistical correlation test was then applied to the same data set. The statistical test adopted is Kendall's correlation test, which is illustrated in more detail in Section 3.6. The value of Kendall's τ (*tau*) can vary between -1 and 1: values around zero indicate absence of correlation, while values tending towards 1 or -1 signal a strong positive or negative association ("as a increases, so does b" vs. "as a increases, b decreases"). Table 3.2 presents the results of this analysis.

	Kendall's τ	***p*-value**
LEXG	0.636	< 0.001
MI	0.006	0.208
LL	0.418	< 0.001

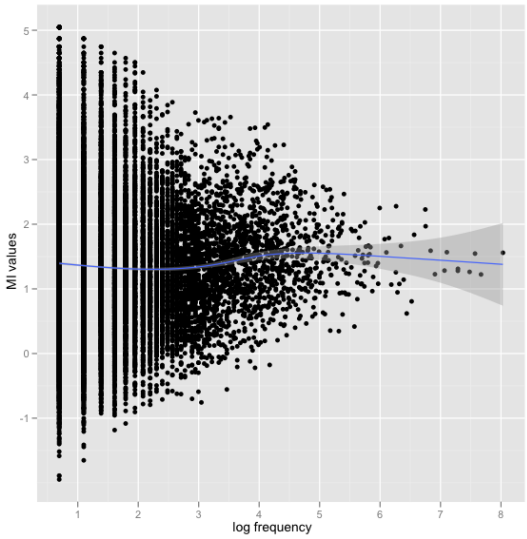
Table 3.2: Kendall's correlation coefficients: AMs $\sim$ frequency

As was expected, LL presents a (moderate) positive correlation with frequency (above significance threshold), while MI does not. The results for LEXG show that it is has a strong positive correlation, which is even higher than in the case of LL.

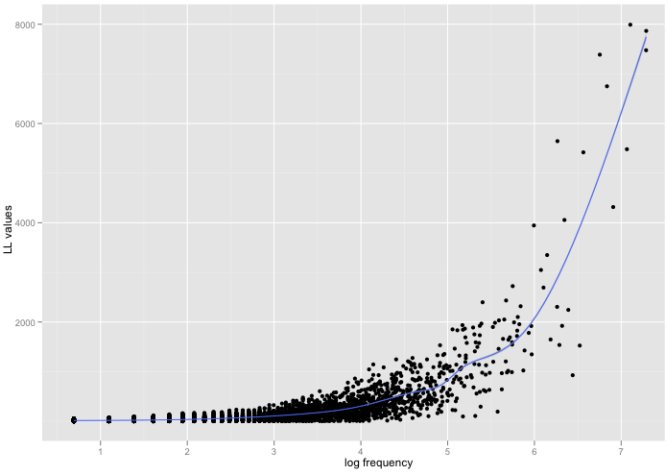
[1] Using logarithmically transformed values is a common procedure when producing plots for skewed distributions, e.g. for word frequency distributions (Baayen 2008:81).



(a) LEXG ∼ (log)FQ



(b) MI ∼ (log)FQ



(c) LL ∼ (log)FQ

Figure 3.3: Correlation between the different AMs and (log)FQ: A-N pairs in UniCoDe-UK.

Based on these observations, it was decided that n-best lists could not be adopted as the only sampling procedure: in order to gain a more comprehensive perspective on the performance of the AMs along the whole frequency spectrum, a mixed sampling procedure was adopted instead. The evaluation set included both the top 10 word pairs for each AM, i.e. the pairs with the highest scores *overall*, and the 10 pairs with the highest association scores in *three frequency ranges* (or strata), i.e. high, medium and low frequencies, defined as the 1^{st}, 2^{nd} and 3^{rd} 33% quantile of the whole frequency range. This is known as a "stratified" sampling strategy.[1] Taking LEXG as an example, the plot in Figure , which is intended for illustrative purposes only, indicates the putative pairs that are extracted with this method, and their distribution along the frequency spectrum.

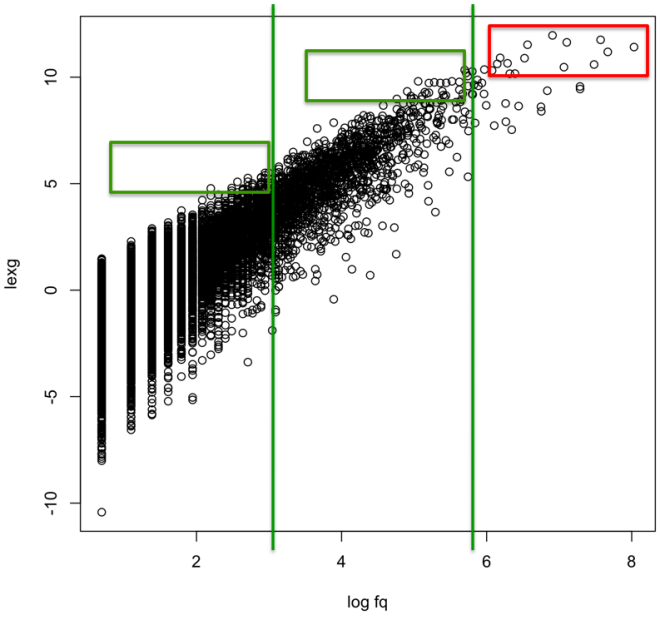


Figure 3.4: Stratified sampling strategy: an example (AM = LEXG).

The red box indicates the pairs with the highest LEXG scores in absolute terms, while the green boxes indicate the pairs with the highest values in the medium and low frequency ranges. In the case of LEXG, the top pairs in absolute terms and the top pairs in the high frequency range coincide, an effect of the strong correlation between the AM and frequency (the same effect is also observed for LL; cf. Section 3.4.3).

[1] The term is used here based on Evert and Krenn (2001); Ellis and Simpson-Vlach (2009) also adopt the same term, though with a different meaning, which was illustrated in Section 2.3.4.2

Table 3.3 presents the whole evaluation set of collocation candidates.

	FQ	LEXG	MI	LL
TOP	wide range	international students	cystic fibrosis	wide range
	first year	more information	connecting uel	more information
	more information	optional modules	stand-up comedy	first year
	final year	financial support	manufactured goods	open days
	third year	wide range	coral reefs	final year
	second year	first year	articular cartilage	optional modules
	optional modules	practical work	reactive compatibilisers	third year
	further information	practical skills	worth two-thirds	second year
	international students	key skills	one-day symposia	typical offer
	transferable skills	subject area	naked eye	transferable skills
HIGH FQ $14 \leq fq \leq 3076$	wide range	international students	smooth transition	wide range
	first year	more information	widest ranges	more information
	more information	optional modules	gross salary	first year
	final year	financial support	linear algebra	open days
	third year	wide range	certified proof	final year
	second year	first year	assessed individual	optional modules
	optional modules	practical work	renewable energy	third year
	further information	practical skills	differential equations	second year
	international students	key skills	partial exemption	typical offer
	transferable skills	subject area	nearest halls	transferable skills
MED. FQ $7 \leq fq \leq 14$	additional tests	historic buildings	reactive compatibilisers	overriding goal
	actual amount	simple notes	nucleic acids	black holes
	active staff	beautiful city	unequalled concentration	strict deadlines
	acceptable subject	architectural practices	liquid bio-fuels	volcanic eruptions
	academic training	departmental website	thermal conversion	rigid deadlines
	work-related learning	video games	volcanic eruptions	manual dexterity
	work-based experience	dynamic region	white man	recommended gcses
	whole spectrum	distinguished scholars	subatomic particles	naval architecture
	weekly timetable	premier venues	manual dexterity	sufficient sketchbooks
	web-based systems	real-life scenarios	proficient enough	premier venues
LOW FQ $5 \leq fq \leq 7$	front line	finished product	cystic fibrosis	coral reefs
	fresh insights	automatic progression	connecting uel	articular cartilage
	french romanticism	design-based competition	stand-up comedy	consultative committees
	french politics	initial concept	manufactured goods	cochlear implants
	french novel	consistent representation	coral reefs	connecting uel
	french law	domestic animals	articular cartilage	bioadhesive polymers
	francophone world	reflexive individuals	worth two-thirds	fast pyrolysis
	francophone country	responsible investment	one-day symposia	stand-up comedy
	foundation-year entry	dramatic text	naked eye	automated dna
	former graduates	serious illness	binding agreement	one-day symposia

Table 3.3: The extracted pairs, ranked by descending AM score.

3.4.3 Final data set

As can be observed in Table 3.3, considerable overlap emerged both among the pairs extracted by different AMs, and among the pairs that the same AM scored as absolute top, as opposed to those that it scored in one of the frequency ranges. It was therefore necessary to identify the repeated items and include them only

once in the final data set. After this processing step, the number of pairs was reduced from 120 to 99.

3.5 Evaluation of collocativity

The 99 pairs forming the final data set were evaluated in three distinct evaluation tasks, each providing a different perspective, or "baseline", on the saliency/collocativity of the word pairs selected by each AM. The tasks and their rationale are described in detail in the Sections 3.5.1 to 3.5.3.

3.5.1 AMs and lexicographic evidence

The first of the baselines against which the AMs were evaluated is lexicographic evidence. Dictionaries, and especially those designed for foreign learners (cf. Moon (2008)), provide valuable benchmarks in two respects: first, they contain a large inventory of phrases whose status as "salient" word sequences was sanctioned by expert lexicographers, and second, they also provide multi-faceted information, e.g. on their typical contexts of use, that can be drawn upon to classify them.

The aim of the task was twofold. First, it aimed at evaluating the "usefulness" of the AMs for an applied purpose, i.e. for the extraction of word pairs that can be considered worthy of dictionary inclusion. Using dictionaries as gold standards (as in Pearce (2002), Daille (1994)), the collocation candidates extracted from corpora by different AMs were classified as "true collocations" if they are present in one or more dictionaries. Here, a similar approach was adopted. The second, inter-related aim was to provide a (loose) classification of the collocation candidates themselves, along two separate dimensions: their degree of "cohesiveness" (distinguishing, e.g. compounds from free combinations), and their degree of specialization (or technicality). There is no attempt at attaining the level of refinement typical of phraseological classifications of collocations (cf. Section 2.3.1), nor to derive generalizations as the types of phraseological units that AMs tend to extract. Rather, this task is meant to provide external evidence against which the results of the other tasks could be compared.

Two distinct dictionaries were used. A collocation dictionary in print form (the *Oxford Collocations Dictionary for students of English* (henceforth OCD, (Lea and Runcie 2002)), and the online version of a learners' dictionary, the *Longman Dictionary of Contemporary English* (henceforth LDOCE, (LDOCE 2003))[1].

The OCD was selected fort two reasons: first, since it focuses exclusively on word combinations, and no space is taken up by definitions, the number of

[1] http://www.ldoceonline.com/ [Last consulted: 15.11.11]

collocations presented (i.e. 150,000, for 9,000 nouns, verbs and adjectives) is likely to be higher than those covered in a general learner's dictionary, which "cannot be expected to contain the same number of collocations as a collocation dictionary. A learners' dictionary addresses multiple needs, its major purpose being to clarify all the senses of a word" (Handl 2008:46). This is a crucial aspect for evaluation purposes: given the low number of collocation candidates tested in this thesis, it was necessary to identify a "gold standard" which included the highest possible number of collocations, to reduce the impact of limited dictionary coverage on the evaluation process. Second, the OCD is designed to list a wide variety of word combinations, "from the strongest and most restricted [collocations], through the slightly less fixed [...] to the fairly open" (Lea and Runcie 2002:821). Based on the British National Corpus, the OCD follows a corpus-driven approach (Tognini-Bonelli 2004) in the choice of the collocations to be included: collocates are not selected adopting phraseological parameters such as, e.g., the distinction between (restricted) collocations or idioms. This reflects the approach to the definition of collocation adopted in this study (cf. 3.2).

Unlike the OCD, the LDOCE presents a more complex *classificatory scheme* of word combinations – although, crucially, no a priori *selection* of collocation candidates is carried out based on this classificatory scheme (cf. Cermák (2006)). The LDOCE was selected as an additional source of lexicographic information both to support the evidence provided by the OCD, and because its classificatory scheme could be drawn upon to draw more refined distinctions among collocation candidates. In particular, the LDOCE presents salient word combinations in three distinct manners: *a)* in bold font, within the entry of a headword: these are the kind of units that dictionary compilers specifically call "collocations", i.e. "the [...] words that are frequently and typically used" with the headword (LDOCE 2003:xviii); *b)* as separate headwords, indicating their status as "compound words [i.e.] groups of two or more words with a fixed form and a special meaning, such as **front man**" (LDOCE 2003:xvi); *c)* within examples, which were "carefully chosen to help show the ways in which a word or phrase is used" (LDOCE 2003:xviii). Furthermore, the online version of the LDOCE (but not its paper counterpart, hence the decision to use the former) presents a "topic categorization" of the headwords, if these are used with a specific meaning in a specialized domain. This information is presented in the format illustrated in Figure 3.5.

Information that was derived from the two dictionaries on each of the 99 collocation candidates was:

- For the **OCD**. Whether the collocation was included in the dictionary or not.
- For the **LDOCE**. If a collocation was included in the dictionary, two variables were taken into account, i.e.:

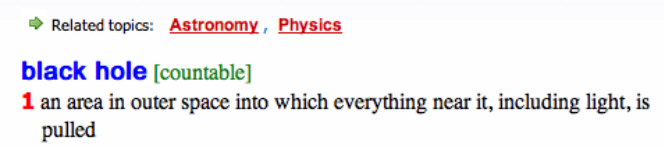


Figure 3.5: Topic categorization in the LDOCE (entry: "black hole").

- the format in which the word combination was presented, i.e. as a separate headword, as a collocation, or as part of an example; in cases where a word combination was listed both under the adjective and under the noun entry, information from the noun entry was retained;
- whether information on topic categorization was provided; if more than one topic categories were provided, the first one was retained;

This information was recorded in a spreadsheet, which formed the basis for the analysis presented in Section 4.2.

3.5.2 AMs and acceptability judgement questionnaires

The second baseline against which AMs/collocation candidates were evaluated was "explicit competence" of expert informants, which was tapped through an acceptability judgement questionnaire (cf. Section 2.3.4.2).

The role of this evaluation task in the present thesis is fundamental, aiming to address the first of the research questions described in Section 3.2, i.e. *a)* whether the output of one or more AMs prompts significantly higher or lower collocativity "ratings" than other AMs, and *b)* whether a set of shared criteria underlying these ratings can be identified, with the ultimate aim of assessing whether consensus emerges as to the salience of a word pair.

The questionnaire was submitted to the participants in the ICAME 32 conference, held in Oslo in June 2011. They were judged to be particularly suitable as informants for the purposes of the present study, insofar as:

- they form a relatively homogeneous community of experts in corpus linguistics, who can be assumed to have a (personal) stance on the notion of salience/collocativity: their judgements could thus be interpreted as a sort of meta-reflection by the corpus linguistics community on the notion itself. Both native and non-native speakers' judgements were considered (information on whether the respondents considered themselves as native speakers of English or not was collected during the experiment). This might be seen as a shortcoming, since it is often assumed that only native speakers

make reliable informants on phraseological issues. Indeed, when non-native speakers' intuitions are tapped, this is usually to test their level of "collocational knowledge", which is implicitly or explicitly assumed to be defective with respect to that of native speakers (cf. Section 2.3.4.2). However, since the aim of this work is to gain a broad understanding of how AMs correlate with salience, and not necessarily with what native speakers judge as salient, including (expert) non-native speakers' judgements in the evaluation might enrich, rather than detract from, the analysis.

- as academics, corpus linguists belong in the same discourse community as the texts' originators 2.2.2.3.1. This might allow them to evaluate collocation candidates in terms of how salient they are with reference to their context of production.

The questionnaire was piloted with a small group of informants comparable to the target population and three different versions were prepared, each presenting the 99 word pairs in the evaluation set in a different random order. In the instructions, participants were asked to consider each word pairs and

> assign each one a score of 1 to 5 according to [their] perception of the degree of lexical association between its members, i.e. according to how strongly, in [their] opinion, the two words are attracted to each other and to what degree they form an **intuitively salient, interesting phrase**. The scores 1 and 2 correspond to low degrees of lexical association ("no or very weak"; "weak"), and 4 and 5 to high degrees of lexical association ("strong"; "very strong"); if you are uncertain about the status of a sequence, assign it a "medium" value of 3.

(The complete version of the instructions is reported in Appendix A). It can be noticed that instructions were vague as to the criteria according to which participants were asked to evaluate the word pairs: no explicit reference is made, e.g., to the notion collocation – the formulation "intuitively salient pair" was used instead. This was done to make sure that respondents would not score down lexically associated pairs simply because they did not fit their (narrower) understanding of the term "collocation".

Space for comments was provided, and comments were encouraged on any of the pairs. Yet, since the task was quite demanding, it was expected that unsolicited comments would not be numerous. For this reason, six pairs were highlighted and comments on these were explicitly requested. These were selected so as to represent different (loosely defined) classes of phraseological units: semantically (non-)compositional sequences (e.g. "final year" vs. "naked eye"),

register-specific ones (e.g. "beautiful city" vs. "open days"), terms (e.g. "cochlear implants") and infrequent modifications of more familiar phrases ("rigid deadlines", related to the more frequent "strict deadlines").

3.5.3 AMs and psycholinguistic data

The last of the baselines was represented by native speakers' "implicit" competence, which was tapped through a Lexical Decision Task (LDT).

The evidence collected in this task aimed to address the second of the research questions outlined in 3.2 above, i.e. assess whether differences emerge in terms of how quickly and accurately the word pairs selected by different AMs are recognized by native speakers of English; this is done with a view to establishing which statistical measure better "reflects" their recognition of word sequences. As was discussed in Section 3.5.3, the accuracy and Reaction Times (RTs) of responses in an LDT are assumed to be directly related to the "representation" of the word pairs themselves in the speakers' mental lexicon: in broad terms, the higher the accuracy level, and the shorter the RT, the more likely it is that a word pair is stored in, and retrieved from, memory as a unit, resulting in faster processing/recognition.

In what follows, the details of the experiment are presented:

- **Participants:** the experiment involved 11 English native speakers, lecturers in linguistic disciplines (e.g. translation and/or linguistics) at the Advanced School for Interpreters and Translators (University of Bologna at Forlì). Six of them are British, two North American, one Irish and one Canadian (average age = 54.3). The target population in the LDT was selected opportunistically, yet an attempt was made to match as closely as possible the one taking part in the acceptability judgement task.

- **Materials:** The 99 word pairs in the evaluation set were used as experimental items. An equal number of "control" items was created: these were obtained by scrambling the adjectives and nouns of the experimental items, so as to form implausible word pairs (the same procedure was adopted by Ellis and Simpson-Vlach (2009)): to count as implausible, the control combinations had to be unattested in two large, general-language corpora of English, i.e. the BNC and ukWaC (Baroni et al. 2009).

- **Procedure:** Participants were tested individually in a quiet room. They were asked to sit in front of the author's laptop, and read the following instructions before the experiment:

  > In this experiment I will show you a series of adjective + noun sequences.
  > I ask you to judge whether you think you are likely to read or hear such

> sequences in English. For example you might think "strong emphasis" or "advanced level" are intuitively plausible word pairs, while "strong level" or "advanced emphasis" are not.
>
> A string is shown mid screen. If you think you are likely to read or hear this in English press "y" on the keyboard; if think you are NOT likely to read or hear this in English, press "n". I am measuring how quickly you do this.

After a practice session in which participants were presented with 20 stimuli not included in the evaluation set, the experiment proper began. The 198 stimuli were presented in random order, and participants had two breaks during the experiment. The RTs and answers to the stimuli were recorded.

- **Hardware and software tools:** the LDT was implemented using the open source *PsyScope X* software (Cohen et al. 2006),[1] and run on a MacBook Pro laptop under Mac OS X Snow Leopard.

certified ranges	premier romanticism	strict subject
assessed buildings	simple cartilage	proficient modules
stand-up DNA	sufficient year	bioadhesive entry
work-related systems	black agreement	foundation-year investment
more concept	front individuals	typical uel
binding practices	key sketchbooks	additional world
French fibrosis	nucleic goal	final bio-fuels
weekly enough	differential deadlines	manual law
French implants	second learning	work-based days
whole students	active salary	widest progression
nearest training	dramatic particles	historic energy
first concentration	francophone polymers	liquid representation
initial halls	worth politics	renewable work
smooth algebra	fast spectrum	consultative experience
responsible novel	beautiful symposia	optional pyrolysis
overriding acids	transferable comedy	web-based year
automated year	recommended dexterity	French transition
acceptable eye	cochlear staff	reactive eruptions
articular city	unequalled website	linear amount
naked information	distinguished offer	serious graduates
cystic information	reflexive line	real-life committees
subject product	gross scholars	partial illness
volcanic support	international reefs	naval deadlines
white year	finished skills	design-based notes
third skills	further timetable	rigid competition
wide tests	manufactured individual	departmental text
practical country	francophone range	open two-thirds
subatomic games	coral skills	academic region
thermal man	automatic equations	connecting scenarios
French proof	former architecture	financial conversion
domestic holes	video compatibilisers	architectural animals
fresh GCSEs	one-day area	dynamic goods

Table 3.4: The control pairs used in the LDT.

[1] Available from http://psy.ck.sissa.it/

3.6 A note on the statistical methods adopted in the analysis of results

This Section provides an overview of the statistical tests that were adopted in the analysis of results. Only a summary description is provided of these tests, mainly with a view to pointing out their relevance for the purposes of this study: as such, it does not purport to present a detailed account of their mathematical and statistical background. This can be found in the work of Baayen (2008), Gries (2009) and Gries (2010b), on which this Section is largely based.

The statistical tests adopted, which were carried out using the open-source software R[1], were the following:

- *Shapiro-Wilk test.* The Shapiro-Wilk test is used to assess whether data (e.g. mean ratings, association scores provided by an AM, etc.) display a normal distribution: if it returns a probability of error (or p) value below 0.05, i.e. the "standard" significance level, this is evidence that the data are *not* normally distributed. Normality in the distribution is a requirement for several statistical techniques: the Shapiro-Wilk test is therefore used as a "pre-test" to decide on the appropriateness of a statistical technique for an analysis (based on whether relevant data turn out to be normally distributed or not).

- *Kendall's correlation test.* This test is used to assess the extent to which two numerical variables are related, or *correlated*: more precisely it tests the behaviour of one variable when the other increases or decreases. It returns a correlation coefficient τ which, quoting Gries (2010b:270), "is close to 1 when there is a strong positive correlation ('the more *a*, the more *b*'), [...] is close to -1 when there is a strong negative correlation ('the more *a*, the less *b*'), and [...] is close to zero in the absence of a correlation"; the test also returns a p value, which indicates whether the correlation found is statistically significant. Following Gries (2010b), Kendall's τ was chosen as a method for correlation analysis since, as we shall see, the Shapiro-Wilk test revealed that in several cases the data under analysis are not normally distributed (cf. 4.3.2.2 below). The more widely used Pearson's correlation coefficient r, which assumes normality of the data, could not therefore be adopted here (Gries 2009:27).

- *Krippendorff's alpha coefficient.* This is one of the so-called measures of *inter-rater reliability*, or *inter-rater agreement*, which are widely used indicators of the degree of consensus among raters (Spooren and Degand 2010):

[1] http://www.r-project.org [Last consulted 21.10.11]

the closer to 1 the coefficient's value, the more consistent are the judgments provided by the raters. Krippendorff's *alpha* coefficient was selected as measure of agreement following Ellis and Simpson-Vlach (2009).

- *Mann-Whitney-Wilcoxon est.* This test (henceforth Mann-Whitney test) is used to verify whether the distributions of two ranked sets of observations are significantly different of each other (when p is < 0.05). As in the case of Kendall's τ, this is a non-parametric test, i.e. it makes no assumptions about the underlying distribution of the data: this makes it particularly suitable to deal with language data, which often violate the assumption of normality (Kilgarriff 2001).

- *Analysis of variance.* Analysis of variance (henceforth ANOVA) is used to compare the means of a dependent variable when it is considered as a function of a multi-level factor (or of multiple factors). By way of example, ratings provided for different word pairs represent a dependent variable, and the AMs that selected them represent a factor with four levels, corresponding to the four AMs themselves: for each level (i.e. AM) mean ratings are calculated, and the significance of the differences of these means is computed. As suggested by Baayen (2008:105-106), when more than two means are compared, running e.g. multiple Mann-Whitney tests may lead to inflated estimates of significance, which of course should be avoided. For each ANOVA that is carried out, two values will be reported and taken into account: and a p value, indicating the significance of the differences found, and a η value (for comparison of two levels) or a *multiple* R^2 value (for comparisons of more than two levels), which, simplifying matters somewhat, indicate the "strength" of the effect of a factor (cf. Baayen (2008:101-108)). If a significant result is returned by an ANOVA, the Tukey Honestly Significant Difference (or TukeyHSD) method is applied, to inspect which level or combination of levels has the highest level of significance.

3.7 Summing up

This chapter has described the methodology underlying the research work conducted for the present thesis. Starting from the statement of the research questions to be addressed, it then moved on to describe the semi-automatic procedure used to construct the corpus, the rationale and parameters used in the establishment of the data set used in the collocation evaluation tasks, the evaluation tasks themselves (collection of lexicographic evidence, acceptability judgement questionnaires, and lexical decision tasks), and the statistical tests used in the

analysis of results. The next chapter, which forms the core of this thesis, reports on and discusses these results.

Chapter 4

Evaluation tasks: results and discussion

4.1 Overview of the Chapter

In the previous Chapter I highlighted the relevance for this thesis of the claim by Wray (2002:66), according to whom formulaic language "is not a single and unified phenomenon", so that "several baselines" are needed to account for what is, or is not, formulaic. I argued that the same applies to the (narrower) notion of collocativity (cf. 2.3.4 and 3.2), and that the evaluation of the output of AMs should be based on various kinds of evidence, focusing on different aspects of collocativity itself.

The term "collocativity" is used throughout the Section in a rather loose sense, and does not refer to a specific notion of collocation as is typical of phraseological approaches (cf. Section 2.3.1), nor does it imply a binary classification of word pairs into collocations and non-collocations (cf. Section 2.3.3). Rather, as was stated in Section 3.2, a word pair is defined as a collocation in this study if it is frequent in a corpus and if "corpus-external" evidence (i.e., lexicographic evidence, explicit endorsement by experts and experimental data) testifies to its salience.

Section 4.2 presents evidence about the lexicographic relevance (or lack thereof) of the expressions in the data set selected for evaluation. Section 4.3 compares the performance of the different AMs in terms of the acceptability judgements obtained from expert informants, providing both quantitative and qualitative insights. Finally Section 4.4 presents the results of the lexical decision task and relates them to the previous two evaluation tasks, showing the connections.

4.2 AMs and lexicographic evidence

4.2.1 Introduction

The first of the "baselines" against which the AMs are evaluated in this Chapter is lexicographic evidence, and namely information on collocations derived from learners' dictionaries, the *Longman Dictionary of Contemporary English* (LDOCE for short), and the *Oxford Collocations Dictionary for students of English* (OCD for short). Bearing in mind the caveats that are connected with the use of lexicographic resources as benchmarks (e.g. their limited coverage with respect to specific specialized discourses and their internal inconsistencies; cf. 3.5.1), this Section pursues a double aim: first, it aims at evaluating the "usefulness" of the AMs in an applied task (simulating lexicographic practice of collocation candidate selection); second, it aims to provide a (loose) classification of the word pairs themselves, along the dimension of their "cohesiveness", and degree of specialization/technicality.[1] This classification will mainly serve as a reference point against which the results of the next experiments may be compared.

A methodological note is in order. As will be remembered from Section 3.4.2, a degree of overlap was observed in the output of the different AMs: 30 pairs were sampled independently for each statistical mesure, resulting in a data set of 120 items, 21 of which were found to be shared by more than one AM. In the present Section, this datum will be largely disregarded in evaluating the AMs' performance: if a word pair was selected by two AMs, e.g. "coral reef" (selected by LL and MI), it will be considered as relevant to both the evaluation of LL *and* MI, but no attempt will be made to assess the performance of the *combination* of these two AMs. The number of pairs pertaining to such combinations is too small to allow generalizations.

The next Section presents the results of the analysis, and Section 4.2.3 summarizes the main findings of the experiment.

4.2.2 Results

This Section presents the results of the analysis focusing on dictionary coverage of the word pairs extracted by different AMs. I will begin by briefly discussing the overall performance of the statistical measures in this setting, and then move on to a more qualitative inspection of the word pairs, with a view to categorizing

[1] The LDOCE distinguishes between strictly "technical" words (and word pairs), and "topic specific" ones (cf. Section 3.5.1): since the latter kind of information was relied on for the classification of word combinations, the term "specialization" and "specialized phrases" will be used instead of their synonyms "technicality/technical phrases", so as not to introduce terminological confusion.

them based on lexicographic evidence.

Table 4.1 presents an overview of the number of word pairs pertaining to each AM that are included in at least one of the dictionaries. As can be observed, the AMs obtain very similar overall results: approximately half of the 30 pairs selected by each of them are found in the dictionaries considered, with numbers varying from a minimum of 14 pairs (for FQ and MI) to a maximum of 16 (LL).

AM	**LDOCE** *and* **OCD**	**%**	**LDOCE** *or* **OCD**	**%**	**Total**	**%**
FQ	4	29	10	71	14	100
LEXG	9	60	6	40	15	100
LL	6	37	10	63	16	100
MI	9	64	5	36	14	100

Table 4.1: Dictionary coverage of the four AMs (number of word pairs).

More substantial differences emerge if results for word pairs included in both dictionaries vs. just one of them are considered. In the case of LEXG and MI, the percentage of pairs found in both the LDOCE and the OCD is around 60%, with the remaining 40% being included in either source; FQ and LL display the opposite trend: the pairs that are included in just one of the dictionaries outnumber by 30/40% those that are found in both of them.

According to the understanding of collocation adopted in this study, which assumes that collocativity has to be assessed against corpus-external evidence, and that *degrees of collocativity* depend on the consensus as to the status of a word pair as a collocation, these results suggest that LEXG and MI outperform FQ and LL in extracting salient collocation candidates: although, overall, the AMs extract a similar number of pairs that may be considered worth including in a dictionary, the former two extract candidates that are more consistently recognized as salient collocations.

This quantitative datum, however, still fails to shed light on the nature of these "salient collocations". To this aim, in what follows an attempt will be made to draw finer distinctions between the collocation candidates, based on the kind of lexicographic evidence described in Section 3.5.1. First, information as to whether the word pairs were presented as separate headwords, as collocations proper within an entry, or as usage examples will be used to classify them on a scale of "cohesiveness" going from maximally cohesive bigrams (i.e. compounds), to "collocation-like" sequences, to free combinations. Second, information derived from the LDOCE as to the typical contexts of usage of word pairs will be used as a criterion to tell apart specialized word pairs and non-specialized ones.

Regarding the first classification attempt, different criteria were used to assign word pairs to one category or another. Starting from the "central" category of

collocation-like sequences, the same criterion used to assess the performance of the AMs was adopted: a word pair is defined as collocation-like only if both the LDOCE and the OCD classify it as such. As was mentioned in Section 3.5.1, in order to assign word sequences to the "peripheral" categories of compounds and free combinations, information was instead derived from the LDOCE only. Two reasons motivated this decision: the first is that the OCD makes no "formal" distinction between different kinds of phraseological units; second, given the focus of the OCD on "the 'slightly less fixed/fairly open' categories" of word combinations (Lea and Runcie 2002:821-822), it could be expected that word pairs belonging in the categories of compounds and free word combinations were less represented in this dictionary (cf. also Lea and Runcie (2002:821-824)). For this reason, word pairs were classified as compounds if the LDOCE lists them as separate headwords (irrespective of whether they are also present in the OCD or not), while word pairs mentioned within examples in the LDOCE were considered free combinations, i.e. combinations in which words are "normally used" (cf. the Introduction to the LDOCE, p. xviii), but which do not qualify as collocations proper.[1] Lacking further evidence, pairs not fitting any of these categories were classified as "other": these correspond to the pairs selected by the OCD only, and to those that were classified by the LDOCE as collocations but that were not found in the OCD.

Table 4.2 presents the results of this classification for each AM. Superscript symbols signal the word pairs that were selected by multiple AMs ("$*$" corresponds to pairs selected by FQ, LEXG and LL, "$\circ$" to the combination of FQ and LL, and "$\triangleright$" to LL and MI): these are repeated in the cells pertaining to each relevant AM (cf. 4.2.1 above). Several observations could be made about the suitability of the classification scheme in characterizing the different word pairs. While, intuitively, the categories of compounds and free combinations seem to describe rather uncontroversially the word pairs they include (e.g. "*black hole*" and "*video games*" as examples of compounds; "*first year*" and "*second year*" as free combinations), the categories of "collocation-like sequences" and "other" are somewhat more problematic. One may wonder, e.g., whether it is appropriate to classify "*further information*" as collocation-like, while "*distinguished scholar*", arguably a more "restricted" sequence (in the sense described by Cowie (1998a)), is classified in the category *other*. By including word pairs in the category *other*, however, I do not mean that they are *not* collocations: rather, that the evidence

[1] A further check on the status of these pairs as compounds or free combinations was performed using the *Cambridge Learner's Dictionary* (http://dictionary.cambridge.org/ [Last consulted 25.11.11]). In all cases, pairs classified as compounds by the LDOCE were also classified as such by the CLD. As for free combinations, information as to their status as less cohesive (and less lexicographically relevant) word pairs was derived indirectly: none of them was present in the CLD.

AM	Compounds	Collocation-like sequences	Free combinations	Other
FQ	front line	further information transferable skills° wide range*	actual amount first year* more information* second year°	academic training final year° francophone country fresh insights optional modules* whole spectrum
LEXG	video games	domestic animals financial support finished product historic buildings practical skills serious illness subject area wide range*	first year* more information*	beautiful city distinguished scholars dramatic text optional modules*
LL	black holes coral reefs▷ open days	manual dexterity▷ stand-up comedy▷ strict deadlines transferable skills° volcanic eruptions▷ wide range*	first year* more information* second year°	consultative committees final year° naval architecture optional modules*
MI	coral reefs▷ cystic fibrosis nucleic acids	binding agreement gross salary manual dexterity▷ naked eye renewable energy smooth transition stand-up comedy▷ volcanic eruptions▷	—	manufactured goods partial exemption white man

Table 4.2: Word pairs classified as compounds, collocation-like sequences, free combinations or "other", split by AM.

for "*distinguished scholar*" being a collocation, derived from a single source, is less stringent than that for "*further information*". Along the same lines, the term "free combinations" does not imply a lack of collocativity of the word pairs it refers to. As with the category *other*, it indicates pairs that are included only in one dictionary (i.e. the LDOCE): in this case, however, a descriptive label could be attached to them, which reflects indications provided by the dictionary compilers.

Adopting a more quantitative perspective, Table 4.3 displays the number of word pairs included in each category for the different AMs. For all but FQ, the category of collocation-like sequences is the most populated, accounting for 37% of the pairs selected by LL, and 53-57% for LEXG and MI. The number of compounds is slightly higher for LL and MI than it is for LEXG and FQ, and

that of free combinations (and of word pairs classified as *other*) is highest for FQ and lowest for MI, with LEXG and LL occupying a middle ground between the other two AMs. Based on the suggested classification, and consistently with the findings discussed above, FQ and MI seem therefore to extract word pairs at opposite ends of the cohesiveness / collocativity continuum: MI tends to select pairs displaying high degrees of cohesiveness (and lexicographic relevance), while the majority of the pairs selected by FQ are either free combinations, or word sequences which are less consistently recognized as collocations by the dictionaries considered. The number of "collocation-like" sequences extracted by LEXG is similar to that of MI, but, compared to it, this AM also extracts a higher number of "less salient" pairs. Finally, LL displays the most "uniform" distribution in terms of the types of word sequences it selects; it should also be noted, in passing, that if the count for compounds is added to that of collocation-like sequences, LL's level of performance is similar to that of LEXG (cf. Table 4.1).

AM	**Compounds**	%	**Collocation-like sequences**	%	**Free combinations**	%	***Other***	%	**Total**	%
FQ	1	7	3	21	4	29	6	43	14	100
LEXG	1	7	8	53	2	13	4	27	15	100
LL	3	19	6	37	3	19	4	25	16	100
MI	3	21	8	57	1	7	2	14	15	100

Table 4.3: Distribution of word pairs according to their status as compounds, collocation-like sequences, free combinations or "other", split by AM.

Moving on, the last dimension along which word pairs will be classified is that of their "topic specialization": given the "composite" nature of the texts making the object of this study, which mix disciplinary/specialized and non-disciplinary topics (cf. 2.2.2), in Section 3.5.1 such dimension was suggested to have special relevance for the description of the AMs' output. In what follows, the word pairs selected by the different AMs will be split into two categories, i.e. specialized and non-specialized word pairs, based on the topic classification provided in the LDOCE; since the OCD does not provide similar information, word pairs included only by this dictionary were excluded from the analysis. Arguably, the degree of topic specialization of a word pair cannot be assessed independently of the contexts in which it is actually used, and establishing a classification based on a single source of evidence increases the risk of making undue generalizations. In order to control for this factor, concordances were checked in the UniCoDe_UK for each word pair which was classified as *specialized*.

Table 4.4 presents the (number of) specialized word pairs found in the output of the different AMs; as in Table 4.2, a superscript "$\triangleright$" indicates those that were selected by both LL and MI (it can be noted, incidentally, that no other com-

AM	Specialized word pairs		Non-specialized word pairs	Total (LDOCE)
FQ	1 (*11%*)	francophone country *[Languages]*	8 (*89%*)	9 (*100%*)
LEXG	2 (*18%*)	video games *[Computers]* finished product *[Industry]*	82 (*87%*)	11 (*100%*)
LL	3 (*25%*)	coral reefs$^{\triangleright}$ *[Biology]* black holes *[Astronomy]* stand-up comedy$^{\triangleright}$ *[Media]*	13 (*75%*)	12 (*100%*)
MI	7 (*58%*)	coral reefs$^{\triangleright}$ *[Biology]* cystic fibrosis *[Illness and Disability]* gross salary *[Finance]* manufactured goods *[Economics]* nucleic acids *[Biology]* renewable energy *[Power]* stand-up comedy$^{\triangleright}$ *[Media]*	5 (*42%*)	12 (*100%*)

Table 4.4: Specialized vs. non-specialized word pairs in LDOCE, split by AM.

bination of AMs was found to yield specialized pairs). The main topic category assigned by the LDOCE to each word pair is indicated in square brackets (cf. Section 3.5.1). As can be observed, the topic categories are extremely varied, and span from clearly defined (disciplinary) fields like "Biology" and "Astronomy" to apparently loose categories like "Languages" and "Power". This heterogeneity was not deemed problematic for the purposes of the present analysis, as long as the specialized topic indicated by the LDOCE at least loosely reflected the contexts of usage in which the word pairs occur in UniCoDe_UK. In all cases, analysis of concordance lines revealed that in fact the LDOCE categories match to a large extent corpus evidence. Table 4.5 displays (selected) concordances for each specialized word and the name of the degree course description from which they were extracted. To mention but some examples, "stand-up comedy" (LDOCE category: *Media*) is found in descriptions of "Creative writing" and "Performance" degree courses, "renewable energy" (LDOCE category: *Power*) in texts on engineering degree courses, and "nucleic acids" (LDOCE category: *Biology*) in texts on "Biomedical sciences" and "Chemistry with Medicinal Chemistry". A single case was observed in which a word pair was not found to be primarily associated with homogeneous topics, i.e. "gross salary", which occurred in descriptions of degree courses in various disciplines: this, however, seems to be primarily due to the word pair occurring in a paragraph which was repeated verbatim across all pages of a single University (i.e. within "boilerplate", cf. Section 3.3).

Once the results of the categorization have been proved to match corpus evidence, quantitative observations can be made more confidently. Going back to Table 4.4, substantial differences emerge between MI and all other AMs: not only is it the measure extracting the largest number of specialized phrases in absolute

Concordance	**Degree course**
This year will be spent in France or a **Francophone country** on a programme of studies in a higher education institution...	*French and Beginners' Russian*
This is complemented by a year spent on a study or work placement in France or other **Francophone country**...	*French with International Studies*
You also work with non-linear, interactive music composition and implementation for **video games** and make use of specialised computer software...	*Contemporary Music Creation*
Today this doesn't just mean computers of course - digital technology is in mobile phones, **video games**, intelligent clothing, electronic music...	*Digital Interaction Design*
On graduation you will be capable of leading and managing creative design projects, from first concept to **finished product**...	*Design and Colour Technology*
Taught by experience staff with proven industrial and commercial design experience, you'll gain a complete knowledge of the creative process from yarn through to **finished product**...	*Fashion Knitwear Design and Knitted Textiles*
The theoretical physics topics covered include quantum field theory and general relativity (which describes cosmology and **black holes**)...	*Theoretical physics and applied maths*
Recent research highlights include the use of ESA's XMM-Newton Space Observatory to study cosmic X-ray sources and the modelling of accretion onto **black holes**...	*Physics with Astrophysics*
Specialist options include writing for children, travel writing (with an opportunity to study abroad), screenwriting, writing for the internet, and **stand-up comedy** ...	*Creative writing*
Students are encouraged to work in formal and informal performance (that is, cabaret, **stand-up comedy**, performance art)...	*Performance*
You'll gain first-hand experience of non-UK ecosystems and their associated fauna and flora, such as **coral reefs**, boreal forests...	*Biology*
Growing public concern over issues such as degradation and destruction of **coral reefs** and tropical rain forests...	*Ecology and Conservation*
You will learn how genetic techniques have become the cornerstone for a host of diverse investigations that include studies of inherited diseases such as **cystic fibrosis**, transmissible diseases...	*Genetics*
The role of molecular genetics in the investigation, diagnosis and design of potential therapies in relation to selected human diseases : e.g. haemaglobinopathies, **cystic fibrosis**, retinitis pigmentosa...	*Health Sciences*
In addition, for 2010 entry UCLan is offering bursaries worth £500 to all UK full time first year undergraduate students, where the principal earner's **gross salary** is less than £60,000 a year.	Miscellaneous
Mechanical engineers are involved at some stage in the conception, design, production, finance and marketing of all **manufactured goods**.	*Mechanical Engineering*
More of our income is spent on services rather than **manufactured goods** and will continue to be so.	*Marketing*
In second year, you will be introduced to the study of proteins, **nucleic acids**, cellular organisation...	*Biomedical sciences*
It involves specialist modules in Medicinal Chemistry where you will study the nature of drug targets e.g. enzymes, receptors and **nucleic acids**...	*Chemistry with Medicinal Chemistry*
This leads to careers in areas as diverse as **renewable energy** systems, power generation, electrical machines...	*Electronic and Electrical Engineering*
New material is also introduced such as composite structures, acoustics, **renewable energy** systems and sustainability, a currently critical area of study for engineers .	*Mechanical Engineering*

Table 4.5: Selected concordances for the specialized word pairs and information on the original context of production (degree course description).

terms, it is also the only measure for which specialized phrases account for the majority of the output (58%). As was the case in the previous analysis focusing on the cohesiveness of word pairs, FQ displays an opposite trend compared to MI, with only one phrase being classifiable as specialized (11%). Finally, LEXG and LL also display similar trends to those which were highlighted in the previous analysis: they tend to select a larger number of specialized phrases compared to FQ, but a lower number of them compared to MI.

4.2.3 Interim summing up

The comparison of the output of different AMs against dictionary coverage has provided indications that, quantitatively, the four AMs extract a similar number of collocation candidates that are worthy dictionary inclusion. Through a more qualitative-oriented analysis, however, we have suggested that AMs extract different *types* of word combinations: FQ seems to give prominence to non-specialized, free/compositional word pairs, while MI displays the opposite trend, highlighting specialized and collocation/compound-like sequences. LEXG and LL occupy a middle ground between the other two AMs, both in terms of the number of cohesive/less cohesive combinations they extract, and in terms of the number of specialized/non-specialized phrases.

As was suggested in the introduction to this Section (4.2.1), given the limited number of word pairs analysed, these results would require confirmation from further studies. However, some clear patterns have emerged, which are consistent with what is known about the behaviour of the different AMs, e.g. MI's tendency to extract term-like (infrequent) sequences, (cf. Baker 2006:112) (cf. Section 2.3.3.1). The classification scheme developed will also provide a benchmark against which the results of the next experiments can be compared.

4.3 AMs and acceptability judgements

4.3.1 Introduction

In Section 3.5.2 I described the aims and rationale of the "acceptability judgment" experiment, consisting in a questionnaire which contained the 99 pairs extracted by the different AMs, and was distributed to the participants at the ICAME 32 conference in Oslo. These were asked to evaluate "the degree of lexical association" between the members of each pair, based on "how strongly [...] the two words are attracted to each other and to what degree they form **an intuitively salient, interesting phrase**"; informants were also encouraged to provide comments motivating their choices. 36 questionnaires were returned (26 by NNSs,

and 10 by NSs).

In this Section the collected data will be analysed, taking into account the following variables:

- information on the sample of word pairs: the 99 adjective-noun pairs themselves, the AM(s) which selected them and the respective scores (i.e. the numerical value assigned to each pair by the AM), as well as the range in the scored list from which they were sampled (cf. Section 3.4.2);
- the informants's responses: the ratings and comments they provided, and information on their level of competence in English, i.e. whether they considered themselves as native or non-native speakers of the language.

This part is structured as follows. In Section 4.3.2 I present a quantitative evaluation of the results, focusing on the ratings provided by the informants and their relation with the AMs: after a brief overview of the pre-processing steps carried out on the collected data (in 4.3.2.1), results are discussed separately for each AM (4.3.2.2), and then compared (4.3.2.3); Sections 4.3.2.4 and 4.3.2.5 conclude the quantitative analysis by investigating the degree to which different variables affected the overall results,namely the sampling method and the informants' level of competence in English. The second half (4.3.3), more qualitative in nature, aims at shedding light on the criteria underlying the ratings in the evaluation task, focusing on different kinds of evidence: the comments provided by the informants (4.3.3.2), the word pairs which obtained the highest and lowest collocativity ratings, and those for which the highest and lowest degrees of consensus among informants was found (4.3.3.3), and, finally, the pairs for which native and non-native speakers provided the most diverging ratings. Section 4.3.4 concludes by summarizing the main findings of the experiment.

4.3.2 Quantitative results

4.3.2.1 Pre-processing of the acceptability judgement data

As will be remembered from Section 3.5.2, informants were instructed to provide ratings on a 1:5 scale, corresponding to a perceived degree of a pair's lexical association ranging from *very weak* (rating = 1), to *very strong* (rating = 5); mid-scale ratings of 3 indicated *uncertainty* as to the lexical association status of the pairs. In order to make judgments more readily intelligible, these were transformed to a scale ranging from -2 to 2 (with 0s corresponding to 3s), so that the "polarity" of the judgment is reflected in the use of negative/positive numbers. If a participant entered a "double" rating (e.g. 3 / 4), after conversion to the new scale (i.e. 0 / 1), the mean of the two values (i.e. 0.5) was retained.

Unless otherwise specified (cf. 4.3.2.5), all analyses are based on ratings averaged across all informants for each pair. Pairs selected by more than one AM are included in the analysis of *all* the relevant measures: an equal number of pairs (i.e. 30; cf. Section 3.4.2) is therefore considered for each AM. While this method disregards a potentially relevant variable, i.e. which/how many AMs selected a specific pair, it was deemed both necessary (in order for the number of "observations" for all the AMs to be equal) and methodologically sound (since the AMs selected these pairs independently of each other). This point is taken up in Section 4.3.2.3.

4.3.2.2 Results split by AM

4.3.2.2.1 Frequency. The cumulative mean rating, averaged across participants, obtained by the pairs selected by FQ is 0.233, the median 0.340 and the standard deviation (henceforth *SD*) 0.646. Figure 4.1 represents graphically the distribution of ratings: this is a "probability density" histogram, where the x axis features the mean rating values and the y axis the probability density values, which can be thought of as the probability of the ratings to occur within the bins on the x axis.

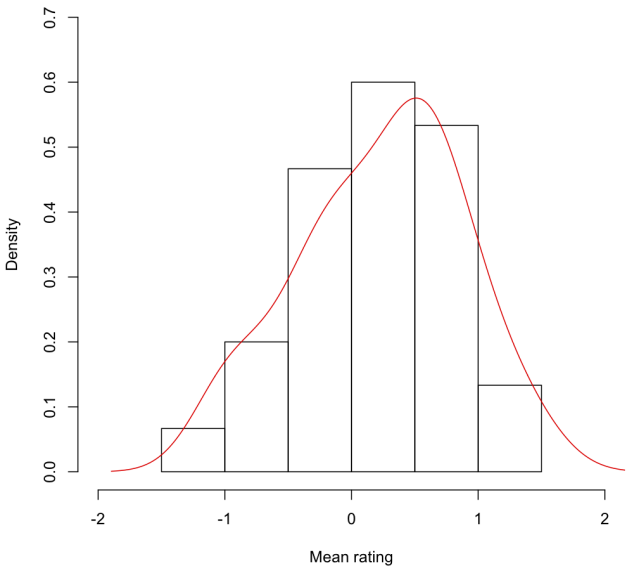


Figure 4.1: Distribution of mean ratings: FQ.

As can be observed, the majority of the ratings cluster around 0 (from -0.5 to 1), and no "extreme" (mean) values (i.e. values tending to -2 or 2) are present.

A Shapiro-Wilk test indicates that the distribution of ratings for FQ does not deviate significantly from normality ($W = 0.9727$; $p = 0.61$), and correlation analysis revealed a very significant positive correlation between ratings and frequency values (Kendall's $\tau = 0.402$, $p = 0.003$).

4.3.2.2.2 Lexical gravity. Pairs selected by LEXG obtained slightly higher average ratings compared to FQ (mean = 0.290; median = 0.375), but they also have a higher standard deviation (SD = 0.823). Figure 4.2 shows that values are more evenly distributed on the x axis, with two peaks around the -1 to 0 and 0.5 to 1 ranges. In this case, too, mean ratings have a normal distribution ($W = 0.962$; $p = 0.349$). Unlike FQ, however, correlation between ratings and LEXG values only approaches significance ($\tau = 0.243$, $p = 0.06$).

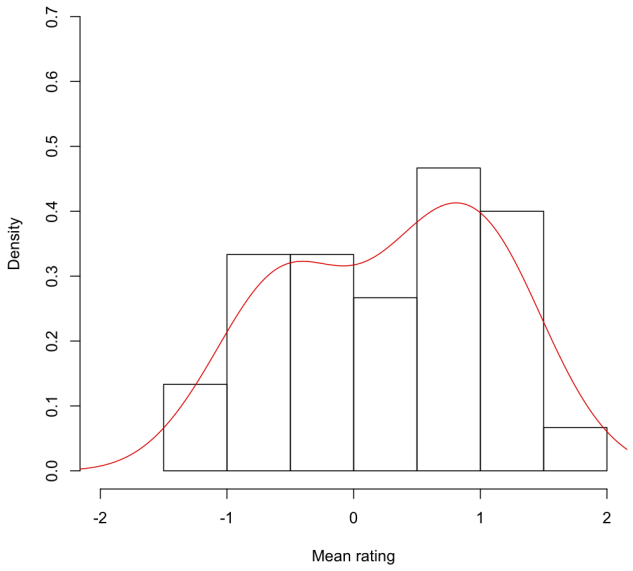


Figure 4.2: Distribution of mean ratings: LEXG.

4.3.2.2.3 Log-likelihood. The average ratings obtained by LL pairs display a distribution which is very similar to that of LEXG: mean and median values (0.287 and 0.354 respectively) are slightly higher than those of FQ, as is SD (0.834); in this case, however, very negative ratings (around -2 value) are found (cf. Figure 4.3). Data are normally distributed ($W = 0.9681$, $p = 0.4874$) and the (positive) correlation between ratings and LL score is significant, although it is less strong than in the case of FQ ($\tau = 0.301$, $p = 0.02$).

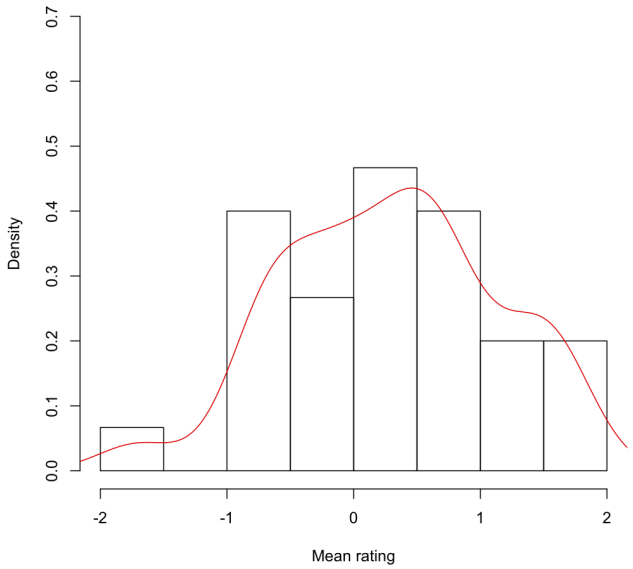


Figure 4.3: Distribution of mean ratings: LL.

4.3.2.2.4 Mutual Information. MI is the measure which obtained the highest average ratings (mean = 0.393; median = 0.646), as well as the highest SD value (0.925). Figure 4.4 shows a clear peak around the 0.5-1.5 range, with other ratings distributed quite evenly across the whole range of values (excluding those around -2, i.e. the lowest range). Unlike all other AMs, the distribution of MI ratings is not normal, with a W value of 0.9286 and $p = 0.04505$, and no significant correlation is found between ratings and MI scores ($\tau = 0.142$; $p = 0.275$).

4.3.2.3 Comparing the results: AMs and acceptability ratings

While in Section 4.3.2.2 I discussed the results of the rating task for each AM in isolation, my aim here is to compare the distribution of ratings across the different AMs, with a view to assessing whether differences can be identified in the degree of (perceived) collocativity of the word pairs they selected.

Yet, before moving on to the comparative analysis of the AMs, a crucial question needs to be addressed. The use of mean ratings as a basis for analysis conceals information on potential discrepancies among the intuitions of different informants: e.g., a mean rating close to 0 could result from most of the informants being uncertain about the degree of lexical association of a word pair (i.e. giving a score of 0), or from a similar number of informants evaluating it as "very strong" and others evaluating it as "very weak". This, of course, is undesirable since one of the aims of the research is to assess whether raters' intuitions were

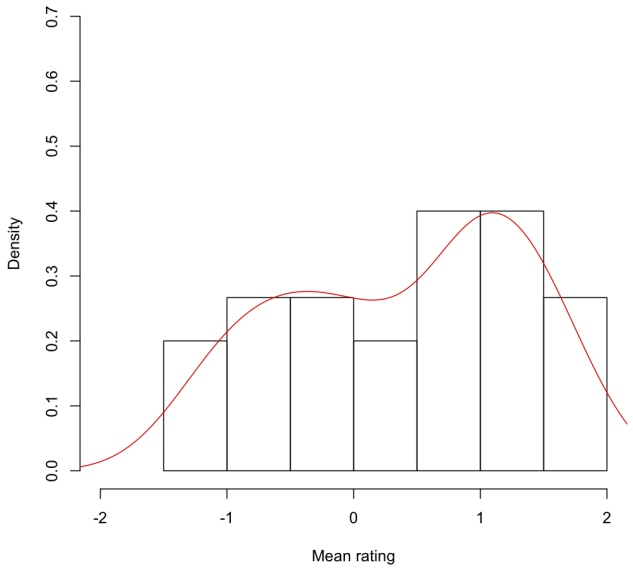


Figure 4.4: Distribution of mean ratings: MI.

consistent with each other, and therefore whether consensus emerges as to how lexical association / collocativity are conceived of (cf. Section 3.5.2).

In order to grasp the extent of agreement among informants, a so-called measure of *inter-rater rater agreement* was calculated (cf. 4.3.2.1 above), which provides an estimate of the consistency with which the 36 informants evaluated the 99 word pairs in the questionnaire. Following Ellis and Simpson-Vlach (2009) and Artstein and Poesio (2008), Krippendorf's *alpha* was selected as a measure of agreement: the resulting value was $\alpha = 0.337$, a level of consensus that Landis and Koch (1977, in Artstein and Poesio 2008:576) define as "fair"; for comparison, the level of agreement is considered as "good" when values are above 0.8 (1977, in Artstein and Poesio 2008:576). Such a low α value would represent an undesirable result in experiments where informants are given precise instructions on how to evaluate experimental items, e.g. in order to produce a gold standard to test the performance of an NLP system (as in Fazly et al. (2007)). I would like to argue that the interpretation of this result in the context of the present experiment should be different. In this case, informants were *intentionally* given vague instructions (cf. 3.5.2), so that they could decide independently on which parameters to adopt in evaluating salience: the low α may therefore be seen as an interesting result in itself, which can be interpreted as a clue to the limited consensus concerning the notion of salience within the community of experts, and one that calls for more careful consideration. In the following Sections, two

variables that are hypothesized to influence the degree of consensus will be investigated, i.e. differences among native and non-native speakers' ratings (in 4.3.2.5 and 4.3.3.4), and discrepancies in the criteria adopted for evaluating salience (4.3.3.2 and 4.3.3.3).

AM	**Mean**	**Median**	**SD**	**Corr. values (AM score/ratings)**	
				τ	p
FQ	0.233	0.340	0.645	0.401	$< 0.01^{**}$
LEXG	0.290	0.375	0.823	0.243	0.06
LL	0.287	0.354	0.833	0.300	$< 0.05^{*}$
MI	0.393	0.646	0.925	0.141	*ns*

Table 4.6: Descriptive statistics and correlation values for the mean ratings of the four AMs.

Postponing discussion of the implications of this finding to Section 4.3.4, I now turn to the analysis of difference among the four AMs considered. The boxplot in Figure 4.5 displays a graphical overview of the distribution of mean ratings for each measure. Each box corresponds to an AM and the black line splitting it in half corresponds to the median of the rating values; the top and bottom edges of the box mark the boundaries of approximately the 50% of ratings around the median, so that the extension of the box along the y axis provides an indication of how scattered the "central" data points are; the dashed vertical lines (called "whiskers") extending above and below the boxes represent the distribution of the remaining 50% of data (if "outliers" are present, these are represented as single dots above or below the limits of the whiskers; cf. (Gries 2009:119)); finally, the "indents" on the right and left sides of the boxes, also called "notches", are used to estimate whether the medians of two different boxes are significantly different (we will return to this below). As was anticipated in Section 4.3.2.2, FQ, and LL have very similar median values (cf. also Table 4.6), suggesting that they provide similarly "good" results in terms of the collocativity of the pairs they select. MI shows a higher median, but also a higher SD (which is reflected in the "height" of its box), pointing at marked differences in terms of its mean ratings: it tends to select pairs which are perceived as more salient compared to other measures, but mean ratings are also more scattered along the range of values, and include very negative ratings (among the 10 pairs with the lowest ratings, 4 were selected by MI; cf. Section 4.3.3.3). On the contrary, FQ and LL have low(er) SD values[1] and hence might be considered as more robust, "reliable" measures, which are likely to extract salient collocation candidates more consistently. Their reliability

[1] The SD value for LL is actually very similar to that of MI and LEXG, but as the boxplot shows, the majority of data points are concentrated in a relatively small area around the median. The value seems to be inflated by the presence of a single negative outlier (the dot near the bottom of the y axis). If this value is excluded, LL's SD becomes 0.646.

is also confirmed by their being the two AMs whose scores display a significant correlation with ratings, which implies that their scores are better able to predict informants' insights as to the collocativity of the pairs they select. LEXG seems to represent a middle ground between the other measures: its average ratings are slightly higher than those of FQ and LL, but so is its SD, which is similar to that of MI.

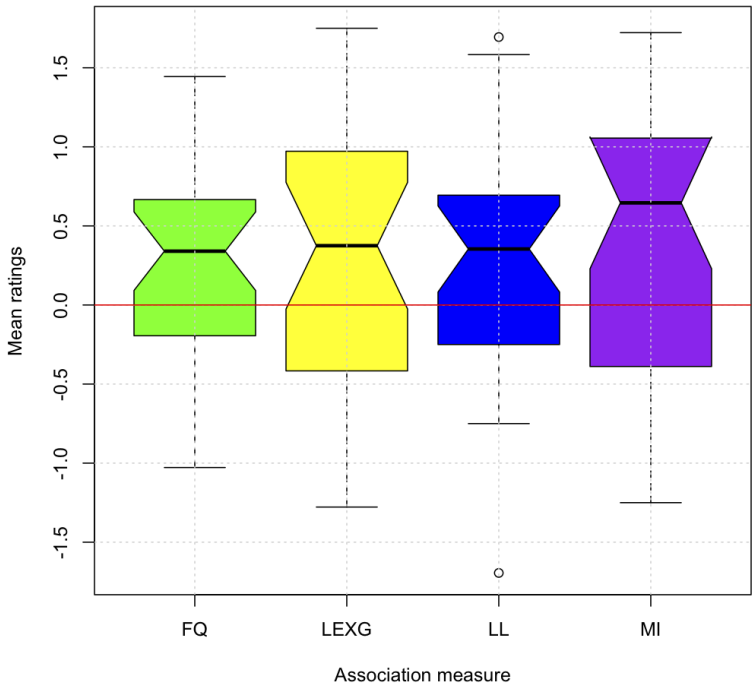


Figure 4.5: Boxplot of the distributions of mean ratings for the four AMs

Inspection of the boxplot representing distribution of collocativity ratings given by expert informants has shown different trends in the performance of the AMs. The question that is asked next is whether, based on these trends, one AM outperforms the others at statistically significant levels. The boxplot provides a preliminary answer to this question: as can be observed, the "notches" of the different boxes are positioned approximately at the same height on the y axis and have a very similar extension; this usually indicates that the differences between the distributions will not be statistically significant (cf. (Gries 2009:119)). This is confirmed by a monofactorial ANOVA (cf. Section 4.3.2.1), aimed at assessing whether the means of the ratings pertaining to one measure are significantly

higher or lower than those of others: with $p = 0.895$, it can be affirmed that no significant difference in the performance of the different AMs emerges ($F_{3,116} = 0.2021$; *adjusted* $R^2 = 0.0052$).[1]

In the next Section, the impact of the sampling method on these results is considered. Before moving on, however, one last test was carried out. For the reasons outlined in Section 4.3.2.1, the analyses presented so far have focused on the pairs that were selected by each AM irrespective of whether they had also been selected by one or more other measures; this, it was argued, is necessary if one is to assess and compare the performance of individual AMs. Yet, information that a degree of overlap emerges in the output of different AMs should not be disregarded. For many practical applications different measures can be "combined" and only word pairs selected by more than one AM taken into account, e.g. to improve the precision of a system for the fully automatic extraction of collocation candidates from a corpus, the rationale being that these pairs are more likely to form salient collocations (an approach adopted, e.g. by Bartsch 2004). To test the hypothesis underlying this approach, a one-tailed *Mann-Whitney* test was performed:[2] the ratings of word pairs selected by multiple AMs (median = 0.667) was compared to those of pairs selected by a single AM (median = 0.111). The former turned out to be significantly higher ($p = 0.03$; $W = 889$). It would therefore seem that word pairs selected by two or more AMs obtain significantly higher collocativity ratings than those selected by just one measure.

4.3.2.4 Other variables: top-scored pairs, frequency ranges

In Section 3.4.2, it was argued that the performance of an AM should be evaluated by taking into account not only the word pairs with the highest absolute scores (i.e. the "top pairs" according to that AM), but also those with the highest scores in *different frequency ranges*. This is the strategy that was adopted in the present study to select collocation candidates: as will be remembered, these consist in the 10 pairs with the highest absolute scores for each AM, and the 10 pairs with the highest scores in the high, medium and low frequency ranges. This makes it possible to compare the performance of the measures in two different ways: one can either focus on their *overall* performance (as was done in Section 4.3.2.3),

[1] When interpreting this result, it should be remembered that pairs selected by more than one AM and their ratings were included in the data sets pertaining to each relevant measure separately. This, of course, is likely to make their distributions more similar than if one considered only pairs that were selected by a single AM. Hence, in order to test whether the analytical procedure adopted had biased the results, a second monofactorial ANOVA was performed, this time excluding the pairs that were selected by more than one AM: the test again returned a non-significant difference ($p = 0.495$; $F_{3,78} = 0.8046$; $R^2 = 0.03002$).

[2] Here, an ANOVA was not necessary, since only two levels of the same variable "number of AMs selecting a pair" were considered.

or adopt a more fine-grained approach and investigate the AMs' performance *a)* when only the *top* pairs are taken into account, and *b)* within each of the three frequency ranges. The latter type of analysis makes the subject of the present Section.

In the first part, I will compare the performance of each AM when either of the two sampling procedures is adopted. The question addressed is whether higher levels of performance are achieved if only top pairs are considered: this is one of the most widespread procedures, also known as the "n-best list" method (cf. Section 2.3.3), for sampling collocation candidates; yet, as has been shown by Evert and Krenn (2001), it is not necessarily adequate for all AMs. I will therefore explore the hypothesis that comparable levels of performance may be obtained by adopting an alternative sampling strategy, i.e. by sampling word pairs taking frequency ranges into account. In the second part of the Section, emphasis will be placed on this second strategy, with a view to shedding light on which measure is better suited at extracting collocation candidates from which frequency range.

In order to assess the performance of the AMs when the n-best list sampling method is adopted, two groups of word pairs are taken into account for each measure, i.e. the ten pairs that it scored as (absolute) top, and those that it scored as top in one of the (high, medium or low) frequency ranges, but that did not make it to the *absolute* top of the lists (which, for this reason, will be called "non-top").

A remark is in order here. For the sake of clarity, so far I have discussed the two groups of pairs as if they were independent of each other: as will be remembered from Section 3.4.2, however, this is not the case, due to the fact that different AMs tend to select pairs belonging in a single frequency range as top candidates. In the data set under consideration, this resulted in the top pairs (henceforth TP) almost completely overlapping with the top pairs in one of the frequency ranges, i.e. the high-frequency range in the case of LEXG and LL (for FQ this is obvious), the low one in the case of MI. On the contrary, the "non-top" pairs (henceforth NTP) coincide with those belonging in the frequency ranges that are not represented in the absolute top of the single measures (i.e. the medium and low ranges for FQ, LEXG and LL; the high and medium ranges for MI).

Figure 4.6 shows the distribution of mean ratings of TPs and NTPs for each AM. As can be observed, FQ and LEXG display similar trends: the median of the ratings (and their means, cf. Table 4.8) pertaining to TPs is higher than that of NTPs, and the latter also display a higher variability, reflected in higher SD values. For LL, the scenario seems slightly different: in this case too, mean and median values are higher for TPs, but the distribution of ratings pertaining to NTPs seems to overlap to a greater extent, compared to FQ and LEXG, with that of TPs (in fact, in the LL plot, the top of the box on the right-hand side

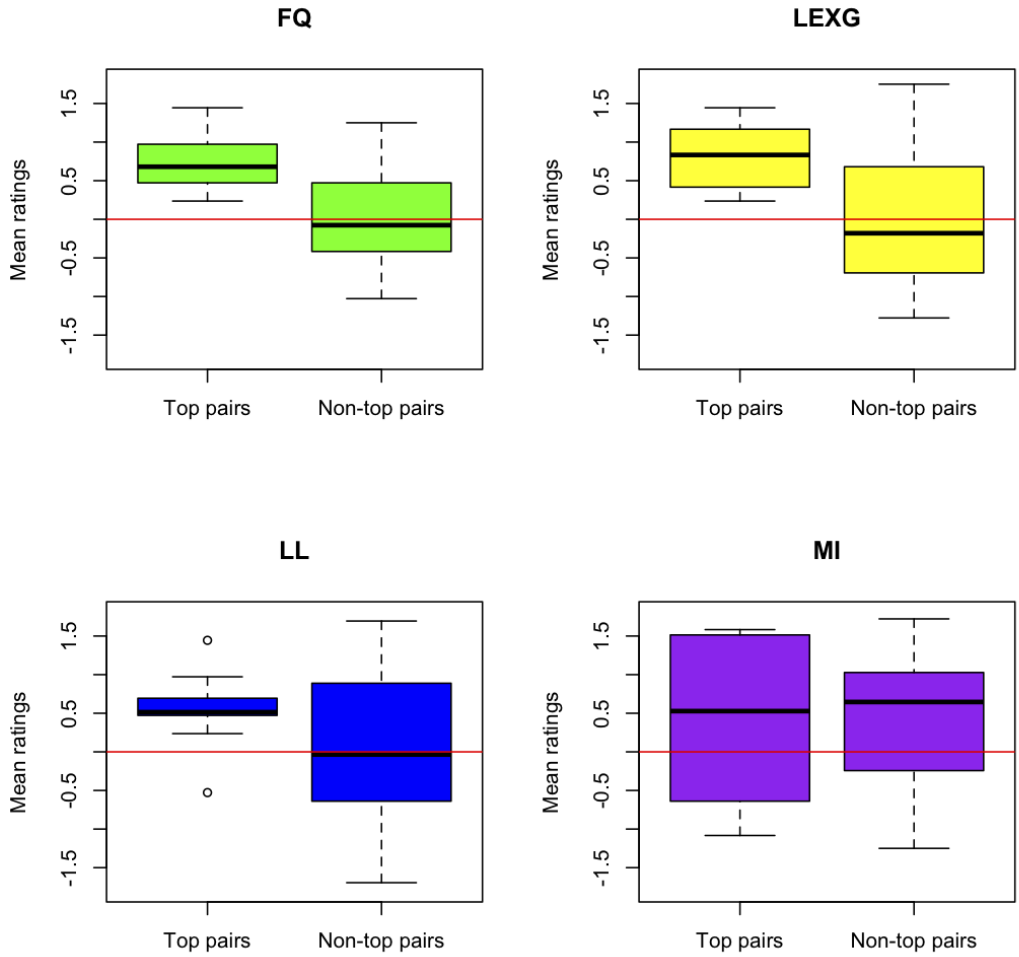


Figure 4.6: Boxplot of the distributions of mean ratings of top-scored and non-top scored pairs, split by AM

is higher than that on the left). Turning to MI, the most surprising results are found: it is the only case in which the median values are higher for NTPs than for TPs (although this is not true of their means), and in which the variability of the former is lower than that of the latter.

As was done in Section 4.3.2.3, these insights were tested for statistical significance. A two-factorial ANOVA was carried out, assessing whether the different combinations of the factors AM and TOP (i.e. whether word pairs were top for one measure or not) resulted in significant differences between mean ratings: a significant correlation was found ($F_{7,112} = 2.262$; $p = 0.0343$; adjusted $R^2 = 0.06911$). Consistently with the results obtained in the previous Section (4.3.2.3),

AM	TOP PAIRS			NON-TOP PAIRS		
	Median	Mean	SD	Median	Mean	SD
FQ	0.680	0.740	0.353	-0.076	-0.020	0.612
LEXG	0.833	0.825	0.416	-0.181	0.023	0.853
LL	0.514	0.546	0.504	-0.035	0.158	0.942
MI	0.528	0.404	1.091	0.646	0.387	0.861

Table 4.7: Descriptive statistics for the AMs: top pairs vs. non-top pairs.

the factor AM alone, with a p value of 0.346, did not contribute to the significance of the difference ($F_{3,112} = 1.115$; $\eta = 0.026$); on the contrary, the factor TOP did contribute significantly ($F_{1,112} = 6.390$; $\eta = 0.050$; $p = 0.013$): this means that that the output of different AMs is evaluated differently depending on whether the top pairs are included or excluded from the test set. In order to assess which specific combinations of AM and TOP had the largest effect on this result, pairwise post-hoc comparisons with Tukey's HSD were carried out (cf. Gries (2009:279)). Differences approaching significance were found for TPs and NTPs selected by FQ (with $p = 0.19$) and between TPs and NTPs selected by LEXG (with $p = 0.14$). The comparisons between TPs and NTPs for LL and MI returned a p value of 1 and 0.9 respectively.[1]

Going back to the hypothesis that was formulated at the beginning of this Section, the results of the analysis suggest that, for FQ and LEXG, the n-best list method for sampling collocation candidates is likely to result in better performance than if word pairs were extracted by "frequency-stratified" sampling, while this is not the case for LL and MI: the performance of these AMs remains stable (or at least does not display significant variation) when either of the two sampling methods is applied. This, it would seem, makes them more appropriate measures when the need arises to extract collocation candidates by applying constraints on the frequency of word pairs.

While crucial for shedding light on the effects that the sampling procedure may have on the performance of the AMs, the analysis just presented fails to consider another important aspect concerning the interrelations between AMs' performance and different frequency ranges. As was also argued in Section 3.4.2, "frequency filters" may be a viable solution to overcome some of the drawbacks associated with the tendency of the measures to select top pairs from a single frequency range. For many applied purposes, this effect is undesirable, since different frequency levels are known to be associated with different types of collo-

[1] While, strictly speaking, p values of 0.19 and 0.14 are not significant, only 3 comparisons, including those just mentioned, returned a p value below the 0.2 level (out of a total of 28 comparisons). The other p value which was below this threshold was obtained for the comparison between TPs selected by LEXG and NTPs selected by FQ.

cation candidates (e.g. high frequency pairs that may be considered as "typical" of the domain under consideration vs. rare but potentially salient ones; cf. Bartsch (2004)). The aim of the analysis that follows is therefore to assess whether certain AMs perform better than others within pre-defined frequency ranges, irrespective of which pairs are scored as top.

In Figure 4.7 the distribution of mean ratings for the four AMs in the three frequency ranges is displayed; descriptive statistics can be found in Table 4.8. Starting from the low frequency range, in the left panel of the boxplot, MI would appear to outperform the other measures in terms of mean and median values (but also in terms of SD, and therefore of "variability" of the ratings), followed by LL. LEXG would seem to provide the worst results, even compared to FQ, displaying the lowest average values and a relatively high SD.[1] In the medium range too, MI seems to be the top-performing measure, with the highest average values and similar levels of variability compared to the other AMs. FQ, LEXG and LL have very similar distributions, with the latter performing slightly worse than the former two, given the high variability of its ratings (cf. the whiskers in the boxplot). FQ, LEXG and LL are instead the best performing measures in the high frequency range, with perhaps a slight edge for LEXG, which obtains the highest mean and median values. Somewhat surprisingly, MI, which is usually believed not to perform well with high frequency words (Evert 2005), also achieves positive average results; its SD value, however, is the highest one.

Visual inspection of the data (and descriptive statistics), suggested that no single measure clearly outperforms the others in the different frequency ranges. MI obtained the highest average ratings when low and medium frequency pairs are taken into account, but it also displayed high values of SD (especially in the low frequency range): this means that ratings tended to be distributed across a wide range of values, an undesirable feature if one wishes to rely on a single AM. The three other measures consistently displayed similar distributions, with LEXG slightly outperforming the other two in the high frequency range, and LL obtaining slightly higher ratings in the low frequency range. FQ, on the other hand, displayed in all cases the lowest SD values.

Three monofactorial ANOVAs were carried out, testing whether the distribution of mean ratings within the three frequency ranges varies as a function of the AMs.[2] In no case differences between AMs were found to be significant (for the

[1] In interpreting this result, it should be borne in mind that the low frequency range includes pairs whose frequency of occurrence is 5 and 7: the paper in which LEXG was introduced (Daudaravičius and Marcinkevičiene 2004) reported that results may be unreliable if LEXG is applied to pairs with frequency < 10.

[2] A single two-factor ANOVA, with frequency ranges and AMs as distinct independent variables, was not deemed appropriate in this case: since I am comparing the performance of the AMs within single frequency ranges, I am not interested in all possible interactions of the

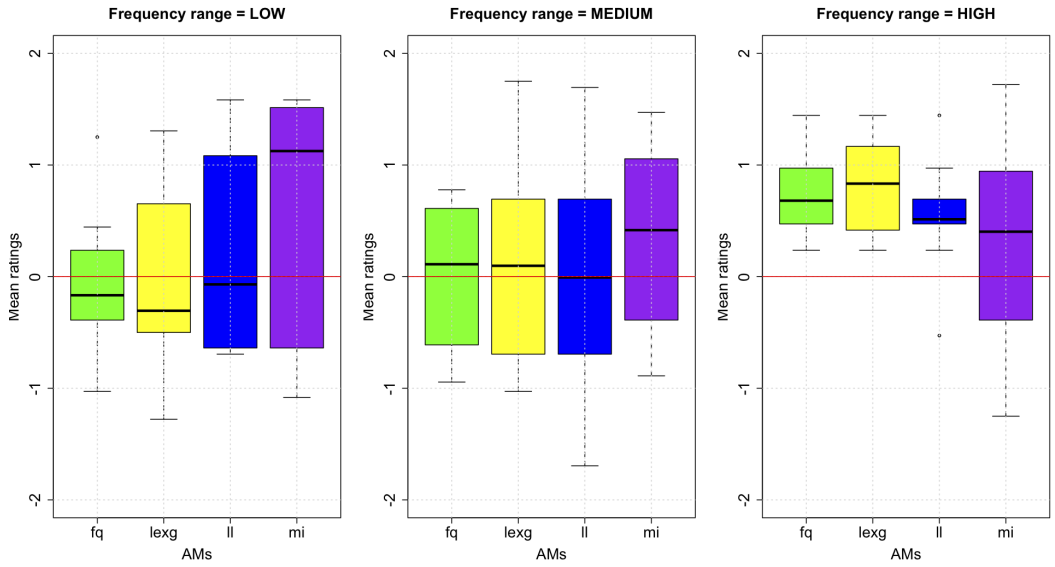


Figure 4.7: Boxplot of the distributions of mean ratings, split by AM, in the three frequency ranges.

AM	**Fq. Range = LOW**			**Fq. Range = MEDIUM**			**Fq. Range = HIGH**		
	Median	Mean	SD	Median	Mean	SD	Median	Mean	SD
FQ	-0.167	-0.049	0.606	0.111	0.008	0.650	0.680	0.740	0.353
LEXG	-0.306	-0.057	0.823	0.097	0.103	0.918	0.833	0.825	0.416
LL	-0.069	0.199	0.882	-0.007	0.117	1.045	0.514	0.546	0.504
MI	1.125	0.582	1.062	0.417	0.340	0.774	0.403	0.257	0.982

Table 4.8: Descriptive statistics for the AMs in the three frequency ranges.

low frequency range: $F_{3,36} = 1.22$; $p = 0.3165$; multiple $R^2 = 0.09228$; for the medium frequency range: $F_{3,36} = 0.2676$; $p = 0.8483$; multiple $R^2 = 0.02181$; for the high frequency range: $F_{3,36} = 0.1226$; $p = 0.1893$; multiple $R^2 = 0.1226$).

A caveat is in order before concluding. Due to the experimental design involving human informants, data samples pertaining to the different variables and their combinations (i.e. AMs, top pairs, and different frequency ranges) were necessarily small, and, as with all small samples, single data points (i.e. ratings for a single word pair) tended to have relatively large effects on the results. Further investigation based on larger samples is therefore required to (dis)confirm the insights this Section has provided. Yet, it is hoped that it succeeded in pointing out

two variables' levels (e.g. comparing LL's performance in the low frequency range vs. MI's performance in the high frequency range).

the need to explore, and the potential benefits of, different sampling strategies when using AMs for extracting collocation candidates.

4.3.2.5 Other variables: native vs. non-native speakers' judgements

In considering collocativity ratings averaged across *all informants*, the analyses in Sections 4.3.2.2 to 4.3.2.4 relied on the assumption that the collected judgments reflected the views of a single, (relatively) homogeneous population. This is consistent with the principal aim of the experiment, i.e. investigating how well the output of different AMs matches intuitions concerning the salience of word pairs, based on insights provided by a community of academics and "experts" in corpus linguistics. Homogeneity was therefore (loosely) determined with regard to disciplinary background and expertise.

In so doing, a potentially influential variable was disregarded, i.e. the informants' level of competence in English. Several studies have pointed out that "collocational knowledge" (Gitsaki 1996), i.e. the ability to correctly produce and recognize collocations, is strictly associated with proficiency in a language (among others: Ellis and Simpson-Vlach (2009); Nesselhauf (2005); Pawley and Syder (1983); cf. also Section 2.3.4). One may therefore legitimately question the appropriateness of conflating results for native and non-native speakers' judgments, and wonder whether this approach is a major source of distortion in the data. It might be hypothesized, e.g., that the lack of consensus that was evidenced in Section 4.3.2.3 actually results from non-native speakers systematically failing to recognize "acceptable" pairs, e.g. due to their low frequency (a hypothesis that was put forward by Gitsaki (1996)). If this were the case, all results obtained so far would have to be reinterpreted in the light of this finding, since the assumption of (relative) homogeneity of the population would not be supported by experimental evidence. In other words, differences in terms of the perceived degree of salience for different AMs might be an effect of raters' language proficiency, rather than of their (personal) informed understanding of collocativity.

Several checks were therefore performed to investigate potential differences and similarities between ratings provided by natives and non-natives (henceforth NSs and NNSs), and to assess whether (and how) these may have influenced the "overall" results discussed in the previous Sections. After recalculating mean rating values for each word pair separating NSs' and NNSs' responses, I first of all tested whether differences emerged between the two groups. Since, as we shall see, this turned out to be the case, I repeated the main analyses that were carried out in Section 4.3.2.3 for NS and NNS data separately, with a view to assessing the effects of these differences on overall results.

Visual inspection of the data (cf. Figure 4.8) reveals that ratings provided by NS and NNS display at least one noticeable difference: the former assigned

	NATIVE			NON-NATIVE		
AM	**Median**	**Mean**	**SD**	**Median**	**Mean**	**SD**
ALL	0.800	0.636	0.925	0.173	0.171	0.789
FQ	0.700	0.597	0.793	0.154	0.093	0.627
LEXG	0.800	0.647	0.855	0.231	0.153	0.838
LL	0.850	0.653	1.006	0.125	0.146	0.799
MI	1.000	0.650	1.065	0.260	0.293	0.896

Table 4.9: Descriptive statistics for the mean ratings of the four AMs, split by type of informant (NS, NNS).

higher collocativity scores than the latter, and did so consistently for all AMs (cf. also Table 4.9). Average NNS' ratings tend to be closer to 0, while mean and median ratings by NS are all above the 0.5 level. Interestingly, however, a degree of similarity also emerges: if one observes how the results split by AM stand in relation to one another within the two groups, in both cases FQ obtains lower average ratings than LEXG (with very similar SD values) and MI obtains the overall highest ratings. Results pertaining to LL would seem to represent an "anomaly" in this scenario: while NSs gave slightly higher scores to LL pairs than to LEXG pairs, NNSs did the opposite. Moreover the degree of variability of the ratings provided by NS is higher than that of NNS: the latter assigned low scores to LL pairs, but did so more consistently than the former (possible motivations underlying these diverging trends will be discussed in Section 4.3.3.4).

Of course, the significance of these differences needs to be tested. In a preliminary step a series of Wilcoxon tests were performed to this aim:[1] results indicated that the medians across NS and NNS groups differed significantly both for aggregated ratings ($V = 6785$, $p < 0.001$), and for ratings split by AM (for FQ: $V = 443$, $p < 0.001$; for LEXG: $V = 448$, $p < 0.001$; for LL: $V = 429$, $p < 0.001$; for MI: $V = 410.5$, $p < 0.001$).

Based on these (significant) results one could hypothesize that NS' and NNS' ratings should have been analysed independently. However, to test this hypothesis, a second, more rigorous analysis is needed. By way of example let us go back to the LEXG / LL difference highlighted in the analysis of the boxplot. The Wilcoxon test indicates that the ratings for both measures are significantly higher for the NS group than for NNS. While interesting in themselves, these results cannot, however, answer the question that is central here: they provide no information as to whether NS assign significantly higher ratings to LL pairs than to LEXG pairs, and whether the same result also emerges in the NNS group.

[1] This test is very similar to the *Mann-Whitney* test used in the previous Sections. However, following (Baayen 2008:83), in this case I set an option specifying that the two groups being compared are dependent, or "paired" (since the mean ratings provided by NS and NNS refer to the same word pairs).

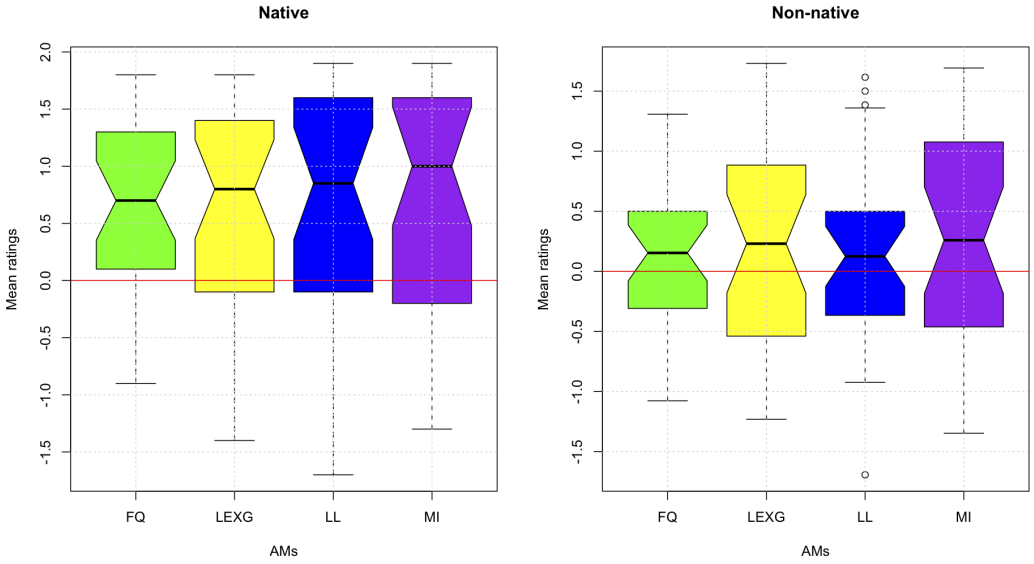


Figure 4.8: Boxplots of the distributions of mean ratings, split by AM, provided by NS vs. NNS.

It might be possible, e.g., that NS are more "sensitive" to LL than to LEXG, and that the contrary is true for NNS, or that these assign similar ratings to LL and LEXG pairs. If such differences were found, then one could reach the conclusion that merging NS' and NNS' ratings has a "blurring" effect on the overall results, and that the fact that no single AM prompts significantly better ratings than another (cf. Section 4.3.2.3) results from this blurring effect.

Accordingly, a multivariate ANOVA (or MANOVA; cf. Baayen (2008:158)) was carried out to test whether mean ratings provided by NS and NNS varied as a function of different AMs. The results of the MANOVA indicated that the effect of the factor AM is non-significant ($F_{3,116} = 0.591$; $p = 0.7375$). Two monofactorial ANOVAs, following the procedure that was adopted in Section 4.3.2.3, were then performed to confirm this result: both in the case of NS and NNS, no AM shows a significant interaction with mean ratings (for NS: $F_{3,116} = 0.0245$; $p = 0.9948$; *adjusted* R^2 = -0.025; for NNS: $F_{3,116} = 0.3465$; $p = 0.7917$; *adjusted* R^2 = -0.01675).

The fact that NSs and NNSs display comparable trends in the ratings pertaining to the different AMs (with the caveat that the former provided *consistently higher* ratings than the latter), is further confirmed by correlation analysis. Table 4.10 shows τ correlation values and and p values for the two groups. The correlation between ratings and AMs' scores has approximately equal strength for NSs

and NNSs for most of the measures: it is intermediately strong and significant for FQ and LL, and non-significant for MI; the trends for both groups reflect therefore those observed for overall results (cf. Section 4.3.2.3). The only case where a slight divergence is found is in the case of LEXG: τ values are very similar for the two groups of informants, but while for NSs correlation is significant ($p = 0.033$), for NNSs it only approaches significance ($p = 0.083$): NNSs' ratings, therefore, have had in this case a slightly "distorting" effect on the overall results, making the correlation for the group of informants as a whole only marginally significant (p for overall results was 0.06; cf. 4.3.2.3 above). Again, I postpone discussion of possible motivations and implications of this finding to Section 4.3.3.4 and 4.3.4.

AM	**NS**		**NNS**	
	τ	p	τ	p
FQ	0.406	0.003	0.364	0.007
LEXG	0.279	0.033	0.223	0.083
LL	0.287	0.027	0.278	0.031
MI	0.111	0.399	0.123	0.343

Table 4.10: Correlation values for the ratings of the four AMs, split by type of informants (NSs vs. NNSs.)

Finally, going back to the question that was asked at the beginning of Section 4.3.2.3, i.e. whether the low degree of consensus among informants was motivated by NSs' and NNSs' ratings being conflated, the answer would seem to be negative. In fact, when *alpha* values are computed for the two groups separately, the difference between the two of them is modest: for NSs, inter-rater agreement is only slightly higher than for NNSs (0.427 vs. 0.305). Based on these observations, it can be concluded with a relative degree of confidence that although NNSs tend to assign significantly lower scores than NSs, their intuitions as to the collocativity of pairs extracted by different AMs do not differ to a major extent.

4.3.3 Qualitative observations

4.3.3.1 Introduction

The analyses in Section 4.3.2 were aimed at shedding light on (quantifiable) differences among the output of the AMs based on the judgments of expert informants. Quantitative differences can be subjected to significance testing, yet they provide no clue as to the motivations underlying the judgements, and hence to the criteria for evaluating salience.

To tap into these criteria, different kinds of evidence are used in this Section. First, an analysis of the comments provided by the raters is carried out, in an attempt to identify and classify the sets of criteria that were adopted in the rating

task (Section 4.3.3.2); these are illustrated with reference to the six word pairs for which comments were prompted (cf. Section 3.5.2). The insights gained from this analysis feed into a second evaluation phase, in which the word pairs with the highest and lowest average ratings, and those with the highest and lowest SD, are inspected, irrespective of the AM that selected them (Section 4.3.3.3). The former should add to the understanding of the criteria underlying high / low degrees of collocativity, while the latter provide clues as to the factors underlying the consensus, or lack thereof, on what a "salient word pair" is in the first place. In Section 4.3.3.4 I conclude by inspecting the pairs for which the most divergence is found between the averaged ratings of NS and NNS, with the aim of looking for possible explanations for the differences highlighted in Section 4.3.3.4.

Before moving on, a note of caution is in order about the use of comments to make generalizations on collocativity criteria. Some of the word pairs that are focused on (namely those in 4.3.3.2 below) were selected *a priori* (cf. Section 3.5.2), so as to stimulate comments on a variety of aspects that were *hypothesized* to underlie salience (e.g. whether a pair forms a compound, a figurative expression, a technical term etc.). Since instructions were intentionally vague about the aspects to be focused upon in the evaluation (differently from, e.g. Evert and Krenn (2001)), the criteria may overlap to a greater or lesser extent (across the whole set of informants and/or with respect to those originally adopted for the *a priori* selection of pairs). Any degree of overlap (or variation) found is seen as an interesting outcome in itself, since it reveals the amount of shared ground within the community of experts.

4.3.3.2 Collocativity criteria: insights from the informants' comments

Based on a thorough scrutiny of all comments, six macro-categories were distinguished, each pointing to a distinct (loose) collocativity criterion. I first illustrate the principles I adopted for assigning comments to different categories, then move on to discuss how these categories (i.e. criteria) relate to collocativity ratings, taking as a case in point the six word pairs that the informants were explicitly asked to comment on.

The six macro-categories of criteria identified are the following (cf. Table 4.11 for examples of comments grouped according to the proposed classification. The whole set of comments, for all word pairs, can be found in Appendix B):

- *Frequency*: this category groups comments in which raters explicitly invoke (intuitively assessed) frequency of co-occurrence as a collocativity criterion (e.g. "*high frequency*", or "*frequent*"; cf. [2] and [4] in Table 4.11). Comments such as, e.g. [1] "*no problem with this – normal*", and [3] "*very common expression*" are also subsumed under this category, based on the

assumption that adjectives like "normal" and "common" refer to the frequency with which the pair is encountered in language.

- *Phraseology*: as the name suggests, this category includes comments which make reference to the various criteria and descriptive labels proposed within "phraseological" approaches to collocation definition (cf. Section 2.3.2.5, 2.3.2.6 and 2.3.1). Two main sub-classes of comments can be distinguished. The first one includes comments mentioning terms and expressions like [5] "*compound*" or [7] "*fixed phrase*", or [10] "*established collocation*", which are likely to be derived from categorizations such as those presented in Cowie (1998b:5).[1] The second sub-class is that of comments alluding to the notion of lexical restriction, which in phraseological approaches is frequently adopted as a criterion for discriminating between "free" and "restricted" word combinations (cf. Section 2.3.2.5). Some examples are: [6] "*'particles' seems like the only word one can have here*", [8] "*what else would stand up go with?*", or [9] "*I can have many different 'worlds'*", which seem to suggest that salience / collocativity depends on the perceived (un)restrictedness in the selection of collocates.

- *Register specificity*: in several cases raters mentioned register[2] as a relevant collocativity criterion. They either suggested a domain in which the word pair may be salient (as in [11] "*management talk – wouldn't say it myself*", or [12] "*medical domain only*"), or by referring to the piece of information, provided in the instructions, that the word pairs in the questionnaire were extracted from "a genre-specific corpus of undergraduate course descriptions"; examples of the latter type of comments are: [13] "*in university context only*", and [14] "*association with university schedules*".

- *Term status*: this category might seem to overlap with the previous one, insofar as (technical) terms are, broadly speaking, vocabulary items that are specific to a certain register / domain. However, a separate category was created since in several cases raters provided comments like [15] "*technical term*", or [16] "*specialist terminology (?) of a different discipline*", without further specifying the domain in which the relevant word pairs are used. This would seem to suggest that technical terms are identified as such regardless of whether informants know the domain to which the terms

[1] Reviewing the work of Gläser (1988), Howarth (1996), and Mel'čuk (1988), the author discusses, among others, the notions of "restricted collocation", "word-like (or semantic) unit" (or "compound", cf. Moon (1998b) in the same volume), and "set phrase".

[2] I use the term "register" here in a rather loose sense, based on the definition by Biber et al.'s (1999:15), according to which "registers are institutionalized varieties or text types within a culture".

themselves belong (and, what is more, regardless of whether they fully grasp their meaning).

- *Lack of familiarity*: this is a "negative" criterion, in the sense that it encompasses cases in which the informants justify their rating by claiming that they do not know the meaning of the word pair in question, or perceive it as unusual. Examples are: [17] "*I don't know what this means without more context*", [18] "*unfamiliar*", and [19] "*what does it mean?*".

- *Other*: the last category groups cases in which comments did not seem to belong to any of the categories just described. It includes: *a)* comments that pointed to criteria adopted by a single rater (e.g. personal experience, as in [21] "*I studied French at university so for me...*", and [22] "*I worked in a university with FYE*", provided by NA_10), or *b)* on a single occasion (e.g. incompleteness of the word pairs [25] "*but with a preposition could be 4/5*"), and finally *c)* comments whose wording is obscure or unclear (e.g. [20] "*possible, but not necessary?*", [23] "*non specific*", [24] "*ugly*", [26] "*style mixture*"), and hinders attempts to infer motivations underlying the judgements.

As with virtually all categorizations, not all cases lend themselves to being assigned to a single category, and several "borderline" comments were found, such as e.g. "*terminology at universities*" (referring to the pair "open days", cf. Appendix B), which can be related to both the "register specificity" and "term status" criteria, or "*common, effectively a compound*" (referring to "cochlear implants", cf. Appendix B), which mixes frequency-related and phraseological observations. This, it is believed, does not undermine the validity of the categorization as a whole.

So far, it will have been noticed, no attempt was made to establish a connection between collocativity criteria and the corresponding ratings in the informants' responses. This is done in what follows by taking into account the "prompted" comments, which are numerous enough to let patterns of association between collocativity criteria and ratings emerge. Comments are discussed according to the word pairs they relate to, and each of these is considered in turn, in alphabetical order. Table 4.12 presents selected comments and the associated ratings for each of the six word pairs under consideration, as well as information on the different raters who provided them (NSs and NNSs); finally, overall mean ratings, averaged across *all* informants, and SD values (cf. 4.3.2.1 above) are reported, so as to provide a more complete picture on the evaluation of each word pair by the participants in the experiment.

- "*Beautiful city*". Two sets of criteria seem to have been relied upon in the evaluation of this pair. On the one hand, frequency criteria are invoked, as

Category	Pair	Comment	Rater info
Frequency	beautiful city	[1] no problem with this – normal	NS_01
	strict deadlines	[2] high frequency	NS_11
	final year	[3] very common expression	NS_14
	responsible investment	[4] frequent, spreading	NNS_28
Phraseology	black holes	[5] isn't that a compound?	NNS_04
	subatomic particles	[6] 'particles' seems like the only word one can have here	NS_10
	cystic fibrosis	[7] a fixed phrase	NS_11
	stand-up comedy	[8] what else would stand-up go with?	NS_11
	francophone world	[9] I can have many different 'worlds'	NNS_28
	serious illness	[10] established collocation	NNS_35
Register specificity	transferable skills	[11] management talk – wouldn't say it myself	NS_01
	articular cartilage	[12] medical domain only	NS_11
	final year	[13] in university context only	NNS_26
	first year	[14] association with university schedules	NNS_35
Term status	nucleic acids	[15] technical term	NNS_28
	subatomic particles	[16] specialist terminology (?) of a different discipline	NNS_35
Lack of familiarity	worth two-thirds	[17] I don't know what this means without more context	NS_01
	optional modules	[18] unfamiliar	NNS_13
	sufficient sketchbooks	[19] what does it mean?	NNS_33
Other	beautiful city	[20] possible, but not necessary?	NNS_04
	French novel	[21] I studied French at university so for me...	NS_10
	foundation-year entry	[22] I worked in a university with FYE	NS_10
	nearest halls	[23] non specific	NNS_28
	wide range	[24] ugly	NNS_28
	more information	[25] but with a preposition could be 4/5	NNS_33
	unequalled concentration	[26] style mixture	NNS_35

Table 4.11: Collocativity criteria: examples of informants' comments (selected)

in examples [2] and [4],[1] (in Table 4.12) and they are usually coupled with positive ratings (i.e. 1). On the other hand, in the cases where phraseological criteria are adopted, ratings tend to be negative (between 0 and -2), based on the consideration that "beautiful" is but one of the adjectives that can co-occur with "city" (e.g. [3], [6]; cf. also [5], where the "acceptability" of the word pair is not perceived as a sufficient criterion to qualify it as a "multiword expression").

- "*Cochlear implants*". A wider variety of criteria have been adopted in this

[1] Actually, example [2] may be considered as a "borderline" case between frequency and phraseological criteria, since it combines the use of the adjective "common" and explicit labelling of the pair as a "collocation". In this case, it was hypothesized that the informant did not refer to collocation in a phraseological sense, but rather to Sinclair's (1996:80) definition of collocation as "frequent co-occurrence of words".

Pair	Comments	Rater ID + rating	Overall mean r. and SD
beautiful city	[1] quite weak, not a collocation [2] common collocation [3] the word city can be combined with several adjectives [4] frequency [5] perfectly acceptable, but not mwe [6] 'beautiful' is one of many possible adjectives	NNS_03: -2 NNS_16: 1 NNS_17: -2 NNS_23: 1 NS_14: -1 NS_30: 0	**Mean:** -0.028 **SD:** 1.521
cochlear implants	[7] specialised term [8] never heard of it [9] there are several types of implants [10] much propagated in medical context [11] ? no clue / maybe a technical term [12] does cochlear go with anything else?	NNS_08: 2 NNS_15: 0 NNS_17: 0 NNS_35: 2 NS_09: 0 NS_11: 2	**Mean:** 1.083 **SD:** 1.131
final year	[13] freq. in academic context [16] in university context only [14] I could have a 'second/third' year [15] I perceive this as a compositional phrase [17] seems typical of uni catalogs [18] very strong in academic context	NNS_04: 1 NNS_26: 0 NNS_28: -1 NNS_36: -1 NS_09: 1 NS_30: 2	**Mean:** 0.972 **SD:** 1.183
naked eye	[19] it's a common expression in a corpus [20] never heard it [21] fully idiomatic: 'to the naked eye' [22] set phrase: meaning not derivable from components [23] hard to imagine this in a course description [24] of course, I'm influenced by Sinclair!	NNS_17: 1 NNS_20: -2 NNS_26: 2 NNS_35: 2 NS_09: -1 NS_10: 2	**Mean:** 1.555 **SD:** 0.908
open days	[25] compound? [26] again, only in academic context [27] unfamiliar to me, meaning not clear [28] doesn't make sense to me [29] compound, effectively – operates as single term [30] conceptually clear referrent with very restricted attributive use of 'open'	NNS_08: 2 NNS_15: 2 NNS_31: -2 NNS_34: -2 NS_14: 1 NS_18: 2	**Mean:** 0.528 **SD:** 1.276
rigid deadlines	[31] not sure, sounds unusual [32] 'rigid' more common than 'strict' with 'deadlines'! [33] 'strict' would sound more natural [34] unfamiliar to me, meaning not clear [35] one of the typical adjectives used with 'deadline' and one of the typical nouns used with 'rigid' [36] I think 'strict' would be stronger than 'rigid'	NNS_03: 0 NNS_16: 0.5 NNS_23: 0 NNS_31: -2 NS_18: 1 NS_30: 0	**Mean:** -0.097 **SD:** 1.308

Table 4.12: Collocativity criteria and ratings: examples of prompted informants' comments (selected).

case. The status of the sequence as a term (in [2]), as an item belonging to the specialized domain of medicine (in [9]), and the lexical restrictions pertaining to the adjective "cochlear" (in [12]) were used as parameters to

assign it a high score (i.e. 2). Interestingly, a case was found where a reverse perspective was taken on the lexical combinatorial properties of its components: instead of focusing on the adjective (as in [12]), one rater focused on the noun (in [9]), and assigned a 0 on the grounds that "implants" can be combined with several adjectives. These opposite views seem to be related to the notion of "directionality", proposed within (mainly phraseological) definitions of collocation (cf. Section 2.3.2.5): based on which part of the pair is considered as the "semantically independent" basis and the "dependent" collocate, the way in which the pair is evaluated changes. Finally, in several cases raters (both NS and NNS) were not familiar with the sequence, and whether or not they recognized it as a term, provided a rating of 0 (in [8] and [11]; cf. also NNS_23 in Appendix B).

- "*Final year*". As was the case with "beautiful city", two main criteria are adopted in the evaluation of this pair. In a few cases, phraseological considerations on semantic compositionality lead to mildly negative ratings (in [14] and [15]). However, the majority of raters seem to focus on the typicality of the sequence in an academic context, and provide positive evaluations (e.g. in [13], [17] and [18]). Only one example was found ([16]) in which an informant recognized "final year" as a typical phrase of university language, and nonetheless provided a "medium" value of 0, possibly suggesting that register specificity does not entail salience.

- "*Naked eye*". Frequency-related and phraseological concerns seem to be the main factors involved in collocativity judgements for this pair: both whether the pair is recognized as a frequent expression (e.g. in [19]), or as an "idiom" or "set phrase" (e.g. in [21] and [22]), ratings tend to be high or very high. Two informants (in [24], and NA_18 in Appendix B) reported that the motivation for their high ratings was to be found in the corpus linguistics literature: they made a reference to the work of John Sinclair (in all likelihood his 1996 paper), who extensively discusses the sequence "naked eye" as an example of "unit of meaning". In one case, a (NNS) rater did not know the phrase, and assigned it a -2. In another case, a negative rating was provided on the grounds that the pair *was not* register specific (in [23]): this is one of the rare examples where register specificity of the pair is associated with a non-positive evaluation (another example can be found in [16] discussed above).

- "*Open days*". Regarding this pair, examples are found in which (mainly NNS) informants assign it a low rating, on the grounds that they do not understand this expression (in [27] and [28]; but cf. also NA_09 in Appendix B, who states that the pair is "not one [s/he] would recognize"). In the

majority of cases, however, "open days" tends to have high scores when it is recognized (by NS and NNS) as a compound-like sequence (e.g. in [25], [29] and [30]), or as a register-specific phrase (e.g. in [26]).

- "*Rigid deadlines*". The main motivations underlying the judgments for this last pair seem to be two, i.e. familiarity with the pair itself, and phraseological concerns. On the one hand, as was the case with all the above pairs, lack of familiarity with the expression resulted in negative to 0 ratings (e.g. in [31] and [34]). On the other hand, phraseological considerations were associated with contrasting intuitions. In some cases, the informants perceived the pair as "typical" (e.g. in [35]); in other cases, the "naturalness" of the pair was called into question (and hence was evaluated more negatively), and the adjective "strict" was suggested as a stronger collocate of "deadlines" (e.g. in [32], [33], [36]).

.

Inspection of how comments relate to ratings proved a valuable approach to tap into the motivations underlying collocativity judgments. Conclusions derived from this analysis are necessarily tentative given the limited number of observations they are based on, yet some clear patterns seem to emerge.

Among the macro-categories of motivations provided by informants to account for their ratings, what I termed "phraseology-related" criteria appeared as the most prominent: for all of the six word pairs, a considerable number of comments was found that could be subsumed under this category. Phraseology-related comments were associated with both positive and negative ratings. If informants recognized a word pair as an instance of a well-established (theory-derived) phraseological category, e.g. a compound ("*open days*"), a set phrase, or an idiom ("*naked eye*"), ratings tended to be high. In several cases, no explicit phraseological "label" was mentioned, but reference was made to criteria such as semantic compositionality ("*beautiful city*") and non-compositionality ("*naked eye*"), or, more frequently, lexical commutability of the components of a word pair: this criterion was either invoked to motivate a high rating ("*cochlear implants*", "*open days*", "*rigid deadlines*"), or to explain a low one, in cases where the word combination was perceived as unrestricted ("*beautiful city*", "*final year*"), or, using Howarth's term (1996:171), as "deviant" ("*rigid deadlines*").

The criteria of frequency and register specificity were found to motivate high ratings in cases where, based on phraseology-related considerations, word pairs instead obtained a low rating: in these cases, the frequency of a pair ("*beautiful city*") or its status as a typical phrase in the academic domain ("*final year*") superseded considerations on its semantic compositionality or commutability of its lexical components, resulting in positive evaluations of its salience. This does

not mean, however, that motivations related to frequency and register-specificity consistently contradicted phraseological insights: depending on the informant, "*naked eye*" was evaluated positively either because it is a (non-compositional) set phrase, or because it is (supposed to be) a frequent expression; similarly, "*open days*" obtained high ratings either on the grounds of its being a compound *or* a register specific expression in the academic context.

Besides lexical restrictions and register specificity, a further criterion was invoked to motivate the collocativity ratings for "*open days*" and "*cochlear implants*", i.e. their being perceived as (technical) terms. Recognition of a pair as a term usually prompted a positive rating. Yet, instances were observed where informants *conjectured* that a word pair was a term and gave it a 0: in these cases the informants' uncertainty might be explained by the fact that they were not familiar with the meaning of the term itself (c.f. comment [11] on "*cochlear implants*" above, but also the comment by NS_19 on "*fast pyrolisis*" in Appendix B).

This reflects a more general trend observed in the data, which involves the last macro-category of comments that was identified, i.e. comments in which informants reported that they did not know the meaning of the word pair to be evaluated. In these cases, what I called "lack of familiarity" consistently resulted in ratings of 0 or below.

4.3.3.3 Degrees of collocativity and consensus: the pairs with the highest / lowest mean ratings and SD

In Section 4.3.3.2, six criteria for evaluating salience were identified through detailed inspection of the comments provided by the informants. The relationship between these criteria and collocativity ratings was then explored, concentrating on the similarities and differences between the views expressed by *single* informants on a small, *selected sample* of word pairs. Here a more wide-ranging perspective is adopted: attention will be focused on mean ratings and the corresponding SD values derived from *all* collocativity judgments. Following a bottom-up approach, I will thus seek confirmation of the patterns identified in the previous Section by inspecting the pairs which were assessed, *overall*, as the most and least salient, and those which prompted the highest and lowest degree of consensus among informants.

Since comments provided by the informants were few and far between for a majority of the word pairs under analysis, and since it is assumed that comments by single informants do not necessarily explain overall patterns emerging from the ratings, the categorization of word pairs provided here will be largely based on external evidence, and namely on the lexicographic categorization that was proposed in Section 4.2.2; corpus evidence will also be tapped to support the claims

made. The aim of this analysis is twofold: first, it aims at assessing whether patterns emerge that point to a shared view (across participants and with reference to lexicographic evidence) of what constitutes a lexically associated, salient phrase, thus taking a further step towards answering question 1 of this study (cf. 3.2); second, it addresses the question whether the motivations that were identified in the previous Section as being central to collocativity judgments are actually able to explain overall patterns emerging from ratings. Admittedly, intuition will play a role in this exploration: conclusions are therefore provisional, and would require confirmation by further studies, in which the relationship between (more precisely defined) phraseological categories and collocativity judgements are investigated systematically.

Highest ratings				**Lowest ratings**			
Pair	**AM**	**Mean r.**	**SD**	**Pair**	**AM**	**Mean r.**	**SD**
video games	lexg	1.750	0.439	unequalled concentration	mi	-0.888	0.887
renewable energy	mi	1.722	0.513	active staff	fq	-0.916	1.079
black holes	ll	1.694	0.709	acceptable subject	fq	-0.944	0.984
coral reefs	ll-mi	1.583	0.692	foundation-year entry	fq	-1.027	1.108
naked eye	mi	1.555	0.908	simple notes	lexg	-1.028	0.941
stand-up comedy	ll-mi	1.514	0.818	assessed individual	mi	-1.083	1.130
volcanic eruptions	ll-mi	1.472	0.774	worth two-thirds	mi	-1.083	1.079
wide range	fq-lexg-ll	1.444	0.734	nearest halls	mi	-1.250	0.937
serious illness	lexg	1.305	0.749	reflexive individuals	lexg	-1.278	0.848
cystic fibrosis	mi	1.305	1.142	sufficient sketchbooks	ll	-1.694	0.668

Table 4.13: Word pairs with the highest and lowest mean ratings.

The left panel of Table 4.13 displays the 10 word pairs that obtained the highest collocativity ratings. If the lexicographic categorization presented in 4.2.2 is adopted to classify these pairs, an interesting scenario emerges: all of the word combinations which were evaluated by informants as the most salient were also included, based on lexicographic evidence, in the categories of "compounds" and "collocation-like sequences". In particular:

- "*video games*", "*black holes*", "*coral reefs*", "*cystic fibrosis*": according to the LDOCE (and the CLD), these word pairs are compounds;
- "*renewable energy*", "*naked eye*", "*stand-up comedy*", "*volcanic eruptions*", "*wide range*", "*serious illness*": these are examples of what we called "collocation-like sequences", i.e. word pairs that were signalled in the LDOCE as collocations, and were also included in the OCD.

Both specialized (e.g. "*black holes*", "*renewable energy*", "*cystic fibrosis*") and non-specialized word pairs (e.g. "*naked eye*", "*wide range*", "*serious illness*") were found among the top-rated collocation candidates, which would seem to suggest that, unlike the degree of "cohesiveness", their degree of specialization was not a major determinant in the evaluation of salience.

The same trends were observed for the word pairs with the *lowest* SD values, i.e. those for which the highest degree of consensus was found (cf. the right panel of Table 4.14). As can be noticed, 7 out of 10 of these pairs overlap with those which obtained the highest ratings, the 3 non-overlapping pairs being:

- "*distinguished scholar*" and "*smooth transition*": these were indicated as collocations by at least one of the dictionaries;
- "sufficient sketchbooks": in this case the low SD value, associated with a very low average collocativity rating, indicates that the word pair was recognized as "non-salient" by the majority of the informants. In fact, concordances reveal that it is an ill-formed word sequence, extracted from boilerplate sections of the texts (in the sentence "photographs are **sufficient sketchbooks** are extremely important" [sic]).

Taken together, these results would seem to point to a high degree of convergence between acceptability judgements and lexicographic relevance: word pairs which were classified, based on corpus-external, lexicographic evidence as "collocation-like" or "compounds" (irrespective of their degree of specialization) were recognized *by the majority* of human informants as *very salient* collocation candidates.

Relying on the same classification scheme, the word pairs for which the *lowest* degree of consensus was found (i.e. those with the highest SD values) display an equally interesting pattern (cf. Table 4.14):

- only one was an example of a collocation-like sequence ("*manual dexterity*");
- five of them were examples of compositional phrases (classified as either free combinations or as "other"): "*beautiful city*", "*first/second year*", "*francophone country*", "*white man*";
- the remaining four pairs were not included in the dictionaries at all.

Focusing on the collocation candidates for which lexicographic evidence is available, a pattern seems to emerge whereby lack of consensus among informants is the most marked in the evaluation of compositional phrases. As we shall see in the next Section (4.3.3.4), a more refined explanation can be put forward for

Highest SD				Lowest SD			
Pair	**AM**	**SD**	**Mean r.**	**Pair**	**AM**	**SD**	**Mean r.**
beautiful city	lexg	1.521	-0.028	volcanic eruptions	ll-mi	0.774	1.472
third year	fq-ll	1.404	0.500	distinguished scholars	lexg	0.774	0.972
second year	fq-ll	1.404	0.472	smooth transition	mi	0.754	0.944
manual dexterity	ll-mi	1.400	0.681	serious illness	lexg	0.749	1.306
first year	fq-lexg-ll	1.394	0.667	wide range	fq-lexg-ll	0.735	1.444
former graduates	fq	1.379	-0.389	black holes	ll	0.710	1.694
proficient enough	mi	1.379	-0.389	coral reefs	ll-mi	0.692	1.583
differential equations	mi	1.376	0.861	sufficient sketchbooks	ll	0.668	-1.694
francophone country	fq	1.374	-0.181	renewable energy	mi	0.513	1.722
white man	mi	1.358	0.611	video games	lexg	0.439	1.750

Table 4.14: Word pairs with the highest and lowest standard deviation.

these divergences, i.e. that such lack of consensus is due to differences between the collocativity ratings provided by NSs and NNSs: the two groups, it will be hypothesized, rely on different criteria for evaluating salience.

Finally, the lowest mean ratings (cf. the right panel of Table 4.13) were associated with:

- word pairs occurring in pages of a single university: "unequalled concentration", "acceptable subject", "foundation-year entry", "simple notes", "reflexive individuals", "worth two-thirds";
- incomplete word pairs: "active staff" (found in the larger phrase "research active staff") , "assessed individual" (⇒ "assessed individual and group reports"), "nearest halls" (⇒ "nearest halls of residence"), "sufficient sketchbooks" (see above).

While having limited relevance for the analysis of the criteria adopted in evaluating salience, these results demonstrate, if anything, that informants' intuition were reliable in identifying "suspicious" collocation candidates.

As a conclusion, and before moving on to analyzing differences between NSs' and NNSs' collocativity ratings, we will (tentatively) address the question whether the criteria that emerged from the analysis of informants' comments are consistent with the overall patterns emerging from the ratings. The answer would seem to be positive. In Section 4.3.3.2 it was argued that the collocation candidates which were perceived as "phraseologically" interesting (e.g. as "restricted collocations

/ compounds / set phrases"), or as "technical terms" tended to obtain high collocativity ratings: the analysis of the top-rated pairs supported this hypothesis. Frequency- and register-related criteria, on the other hand, would seem to play a marginal role in the evaluation of salience: for instance register-specific pairs like "first/second/third year" obtained ratings close to 0 (but a high SD value, a point we will return to in the next Section).

4.3.3.4 NS vs. NNS: the word pairs with the most diverging ratings

Section 4.3.2.5 presented the results of a quantitative comparison of collocativity ratings provided by NSs and NNSs, which revealed that, overall, the intuitions of the two groups as to the salience of word pairs do not differ significantly. In this Section, a more qualitative-oriented analysis is carried out: focusing on the collocativity ratings assigned to *individual* word pairs – rather than to *sets* of word pairs selected by the same AM –, we will address the question whether different "types" of collocation candidates prompt diverging results in the two groups of informants. This is done by calculating, for each word pair, the difference between the ratings provided by NS and NNS and inspecting the pairs for which this difference shows the highest values.

Native					**Non-native**				
Pair	**AM**	**M. NS**	**M. NNS**	**Diff.**	**Pair**	**AM**	**M. NNS**	**M. NS**	**Diff.**
white man	mi	1.700	0.192	1.508	automated DNA	ll	-0.500	-1.000	0.500
second year	fq-ll	1.500	0.077	1.423	linear algebra	mi	1.115	0.700	0.415
departmental website	lexg	1.200	-0.154	1.354	recommended gcses	ll	-0.654	-1.000	0.346
first year	fq-lexg-ll	1.600	0.308	1.292	reactive compatibilisers	mi	-0.462	-0.800	0.338
manual dexterity	ll-mi	1.600	0.327	1.273	worth two-thirds	mi	-1.000	-1.300	0.300
beautiful city	lexg	0.800	-0.346	1.146	domestic animals	lexg	1.192	0.900	0.292
french politics	fq	0.600	-0.538	1.138	former graduates	fq	-0.308	-0.600	0.292
third year	fq-ll	1.300	0.192	1.108	additional tests	fq	0.154	-0.100	0.254
french novel	fq	0.800	-0.308	1.108	naval architecture	ll	-0.154	-0.400	0.246
fresh insights	fq	1.200	0.154	1.046	reflexive individuals	lexg	-1.231	-1.400	0.169

Table 4.15: Word pairs with the greatest difference in mean ratings between NS and NNS.

Table 4.15 displays the results of this analysis. The right panel presents the

collocation candidates to which NS assigned a higher rating compared to NNS. In all cases, these word pairs were evaluated as salient by NS (with ratings in the 0.6/1.7 range), while NNS evaluated them as relatively non-salient (range: -0.5/0.3). As was mentioned in the previous Section, most of the word pairs included here *a)* largely overlap with those for which the highest SD values were observed, and *b)* are compositional phrases (cf. the classification provided in 4.3.3.3 above, but also, e.g. "French novel", "French politics", "third year").

Inspection of the word pairs that NNS scored higher than NS (presented in the left panel of Table 4.15) reveals a completely different scenario: in most cases, these word pairs were given ratings tending to 0 by NNS, while NS assigned more "extreme" negative scores. By way of example, "automated DNA" was given a score of -0.5 by NNS, and -1 by NS. Only three exceptions were observed, i.e. "linear algebra", "domestic animals" and "additional tests".

This analysis would seem to suggest that NS judge positively pairs that are plausible, irrespective of whether they are "phraseologically interesting", and that NNS are generally more "cautious" in their ratings. The latter behaviour seems akin to what Pym (2008) has called "risk aversion", i.e. the tendency for translators to opt for low-risk options whenever possible. Like translators, NNS respondents in this task also seem to be risk averse.

4.3.4 Interim summing up

An acceptability judgement task was set up to compare the performance of different AMs in predicting experts' intuitions on the salience of word combinations, with a view to answering research question 1, i.e. "Do AMs predict experts' intuitions on the salience of a collocation? And if so, does a given AM predict them better than the others?" (cf. 3.2). Based on the analysis, this question seems to require a two-sided answer. Quantitative observations suggested that no AM clearly outperformed the others in this task: FQ, and the AMs most highly correlated with it, i.e. LL and (to a lesser extent) LEXG, were able to predict collocativity ratings, while MI was not; on the other hand, however, the word pairs extracted by the latter measure obtained, on average *higher* ratings. This, it was argued, is due to the fact that MI, unlike FQ and LL, is not a "robust" measure: it gave prominence to word combinations that were perceived by participants as very salient, but it also extracted a higher proportion, compared to the other two measures, of pairs that were perceived as non-salient (cf. its high overall SD value). The word combinations extracted by LEXG obtained slightly higher ratings than those of FQ and LL, but in this case, too, a relatively high SD value was observed.

The influence of the stratified sampling strategy on AMs' performance was analyzed. This revealed that, while stratified sampling *negatively* affected the

performance of FQ and LEXG, it had no such effect on LL and MI. Interestingly, MI performed (slightly) better in the high/medium frequency ranges than it did in the low one.

The main rationale underlying the separate analysis of NS and NNS comments was also primarily methodological, aiming to confirm homogeneity of the two groups, or, in other words, to assess the extent to which results would have been different if only one category of informants had been taken into account. Yet this analysis also revealed differences between the two groups that seem interesting in their own right, such as the tendency for NNS to avoid extreme judgements, possibly as a result of risk aversion, and the tendency of NS to consistently score compositional/free combinations higher than NNSs.

Turning to the second part of research question 1, i.e. "whether consensus emerges as to what constitutes a salient collocation", quantitative results suggested that a *low* degree of agreement emerged from the collocativity ratings. However, a more qualitative-oriented analysis of individual word pairs (rather than sets of word pairs grouped according to the AM that selected them) and of informants' comments, revealed some interesting patterns. Overall, the main criteria employed in assessing salience were found to be "phraseological": this was evident from the correspondence between high ratings/low standard deviations (meaning a *high* degree of consensus), with lexicographic categories such as compounds, terms and restricted-collocation-like units. Comments confirmed that lexical substitutability played a crucial role in evaluation. This observation has to be taken with a grain of salt, however, since *a)* other criteria were underrepresented in the data set (e.g. there was a single pair for which the criterion of semantic transparency was relevant, namely "naked eye"), and *b)* no one-to-one correspondence was found between a word pair and the criteria that were used to evaluate salience – which, arguably, also explains the overall low degree of consensus. The subjective nature of informants' intuition was particularly evident in comments that centered on phraseological parameters: e.g. in the case of "rigid deadlines" intuitions about the "naturalness" of the adjective in the word sequence were discordant, and in the case of "cochlear implants", depending on which of the noun or the adjective was considered as a *base* or as a *collocator* (adopting Hausmann and Blumenthal's (2006) terms; cf. 2.3.1), conflicting ratings were provided.

4.4 AMs and psycholinguistic data

4.4.1 Introduction

The experiment presented in the previous Section foregrounded the role of experts' intuition as a touchstone for evaluating the performance of different AMs. Adopting a bottom-up approach, I aimed at exploring the notion of collocativity as it emerges from explicit, theoretically informed judgements on the lexical association/salience of word pairs. The present Section describes the results of the second evaluation task (cf. 3.5.3), in which *psychological* salience is adopted as a baseline for comparing the AMs.

A lexical decision task (LDT for short) was set up for this purpose, involving 11 English NS, lecturers at the Advanced School for Interpreters and Translators (University of Bologna at Forlì). Using the PsyScope software (Cohen et al. 2006), they were presented with 198 adjective-noun sequences on the author's laptop screen (i.e. the 99 experimental pairs and an equal number of control items), and were asked to press either "y" or "n" on the keyboard according to whether they thought the word sequences were plausible in English. Their answers and reaction times (RTs) were recorded.

By assessing how these two variables correlate with the output of different AMs, the present Section aims at answering research question 2 (3.2), i.e. whether differences emerge in terms of how quickly and accurately the word pairs selected by different AMs are recognized by NSs; this is done with a view to establishing which statistical measure better "reflects" NSs' processes of recognition of word sequences. The second, interrelated aim of this Section (cf. research question 3) is to compare the trends emerging from implicit, indirect clues pointing to psycholinguistic salience and those emerging from explicit collocativity judgements (4.3), in order to assess the extent to which different baselines support or contradict each other in defining collocativity.

For ease of presentation, the structure of the present Section mirrors closely that of Section 4.3: after describing the pre-processing steps performed on the data collected in the LDT (4.4.2), results pertaining to the accuracy and RTs obtained by each AM will be first discussed separately (4.4.3.1), and then compared across different AMs (4.4.3.2). Subsection 4.4.4 is devoted to more qualitative observations: drawing on the lexicographic categorization proposed in Section 4.2, the pairs which were associated with the shortest RTs will be inspected with a view to explaining the diverging trends displayed by the AMs in the LDT. Finally, Subsection 4.4.5 brings together collocativity judgements and psycholinguistic data to provide a fuller picture on the evaluation of the AMs emerging from this study.

4.4.2 Pre-processing of the LDT data

This Subsection describes the pre-processing steps that were performed on the LDT data in order to obtain the final data set on which the analyses in the following Subsections are based.

The "yes/no" answers provided by each participant for the 198 stimuli and the RT values were imported into a spreadsheet. The number of "yes/no" answers was counted for each experimental pair to provide a measure of the accuracy with which it was recognized as plausible by participants. As for RTs, following standard practice (cf., e.g. Ellis and Simpson-Vlach (2009:67) and Durrant and Doherty (2010:110)), reaction times of less than 200 *milliseconds* (henceforth *ms*) or more than 3 standard deviations above the mean for each participant were replaced with the mean for that participant. The former were likely to be cases in which respondents pressed the y/n key by mistake before they could actually read the word pair on screen, while the latter could result from distractions during the experimental session (e.g. one participant reported being distracted by a noise from outside the room). These data points made up 1.65% of all responses. Five cases (0.2% of all responses) were found in which respondents provided two answers for the same word pair: this was due to the fact that they pressed the wrong key (e.g. they pressed t or m on the keyboard instead of y/n): in these cases, the second answer and the RT of the first one were retained, the rationale being that recognition of the word pair occurred when the first key was pressed, and that the answer participants wanted to provide was the second one.

For experimental pairs, RT values were averaged across all participants based on the RTs of correct y answers; RTs of wrong answers, i.e. cases in which respondents did not recognize an experimental pair as being plausible, were discarded. The RTs of 5 control pairs were also discarded altogether: since all of the participants recognized them as plausible, these pairs made poor controls, despite their not being attested in two very large corpora. Investigating the possible reasons behind the plausibility of these unattested sequences – *consistent insights*, *actual exemption*, *practical venues*, *work-related systems*, *binding practices* – would make an interesting side-project in its own right, that could not be pursued here for reasons of space. The list of control pairs can be found in Table 3.4.

Finally, the averaged RTs of experimental pairs were preliminarily inspected by means of a boxplot (cf. Figure 4.9). This showed 4 marked outliers, i.e. the dots appearing above the upper limit of the dashed line, one for LEXG and MI, and two for LL. Inspection of these pairs – *reflexive individuals* (LEXG), *fast pyrolisis* (LL) and *connecting uel* (LL and MI) – in UniCoDe_UK revealed that they were "suspicious" collocation candidates, since they occurred in sentences which were repeated verbatim across different pages of the same University. While the presence of these pairs in the AMs' scored lists might be relevant in their

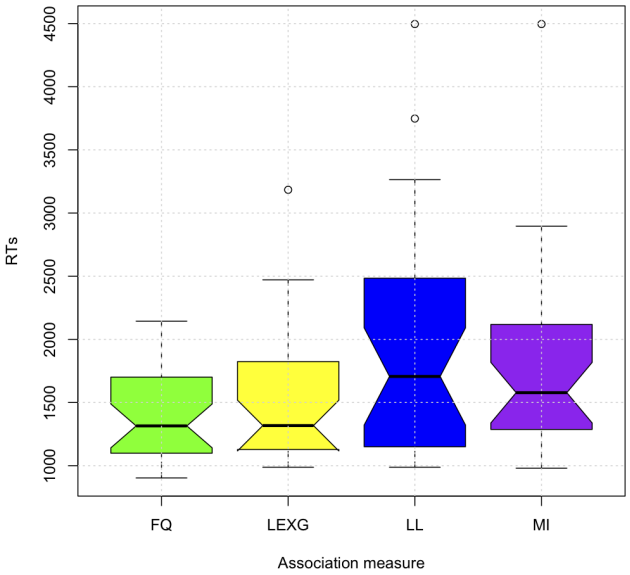


Figure 4.9: Boxplot of the RTs associated with the four AMs (preliminary).

evaluation, e.g. to assess the extent to which AMs are sensitive to noise in the corpus, it was decided to exclude them from the analysis for methodological reasons, i.e. because outliers "might dominate the outcome, partially or even completely obscuring the main trends characterizing the majority of data points" (Baayen 2008:31). Moreover, these pairs obtained low accuracy scores (i.e. they were mostly recognized as implausible), and the averaged RTs would therefore reflect the decisions of a minority of the participants (e.g. *connecting uel* was recognized as plausible by only 2 out of 11 participants).

4.4.3 Quantitative results

4.4.3.1 Results split by AM

In Subsections 4.4.3.1.1-4.4.3.1.4 accuracy scores, defined as the percentage of correct "yes" answers out of the total number of answers provided (pertaining to individual AMs), and descriptive statistics on the distribution of RTs are presented for each AM separately. A comparison between the RTs for experimental and control pairs is then carried out: this makes it possible to evaluate whether experimental pairs are indeed associated with significantly faster processing/recognition times than implausible ones, which is assumed to testify to

their psychological salience. As was done in Subsections 4.3.2.2.1-4.3.2.2.4, correlation values between the AMs' scores and RTs are then calculated, in order to assess the degree to which different AMs are able to predict NSs' RTs.

4.4.3.1.1 Frequency. The accuracy with which pairs selected by FQ were recognized as plausible by the participants in the experiment was very high, reaching 98%. The averaged RTs of the pairs selected by FQ had a mean value of 1416 ms, a median of 1315 ms and a SD of 381.5 (cf. also Table 4.16). As can be observed in the left panel of Figure 4.10, the majority of RTs cluster in the range of 1000-1500 ms, and no values above 2500 ms are found. A Shapiro-Wilk test reveals that RTs do not display a normal distribution ($p = 0.012$, $W = 0.9067$), and hence non-parametric tests will have to be adopted for their statistical analysis.

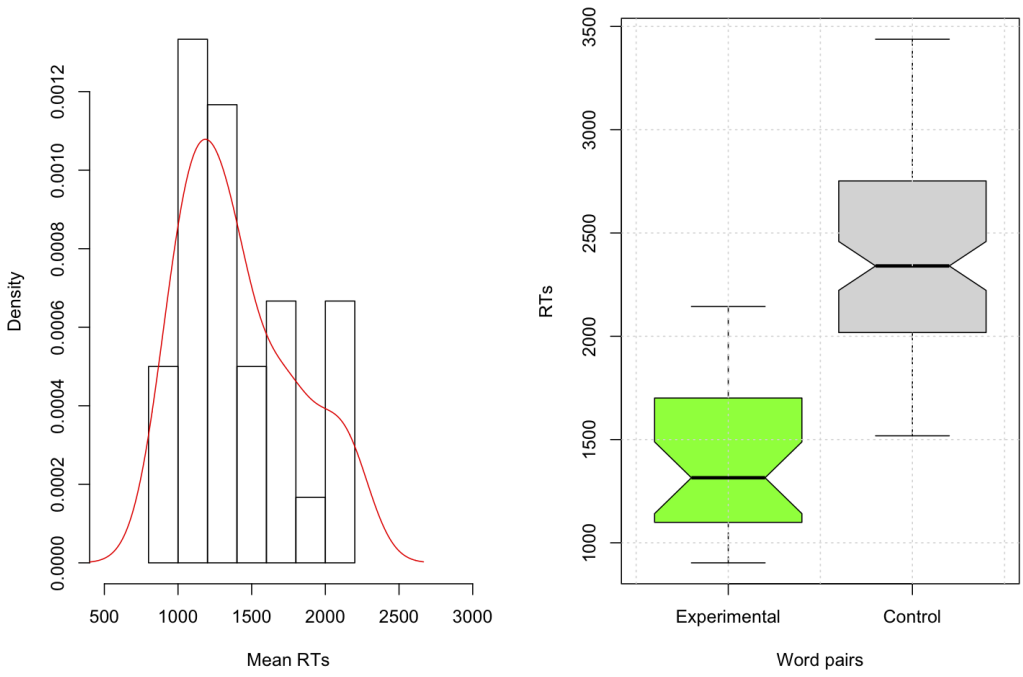


Figure 4.10: RTs and comparison of experimental vs. control word pairs: FQ.

The right panel of Figure 4.10 provides a graphical overview of RTs of experimental pairs and those of control pairs (median = 2340, SD = 441.4). A Mann-Whitney test comparing the RT values associated with the two conditions

indicates that the former are significantly lower than the latter ($p < 0.001$; $W = 154$).

Finally, correlation analysis points to a relatively weak but significant *negative* correlation between FQ values and RTs (τ = -0.264, $p = 0.052$): the higher the FQ value of the word pairs, the shorter the RTs (and hence the more "psychologically salient" the word pairs).

4.4.3.1.2 Lexical gravity. The accuracy pertaining to LEXG word pairs was only slightly lower than that of FQ (95%). The median of averaged RT values was 1316 ms, their mean 1484 ms and SD 460.2. The distribution of RT values is also very similar to that of FQ, with the majority of RTs falling in the 1000-1500 ms range (cf. Figure 4.11, left panel). In this case, too, data are not normally distributed ($p < 0.001$, $W = 0.8554$).

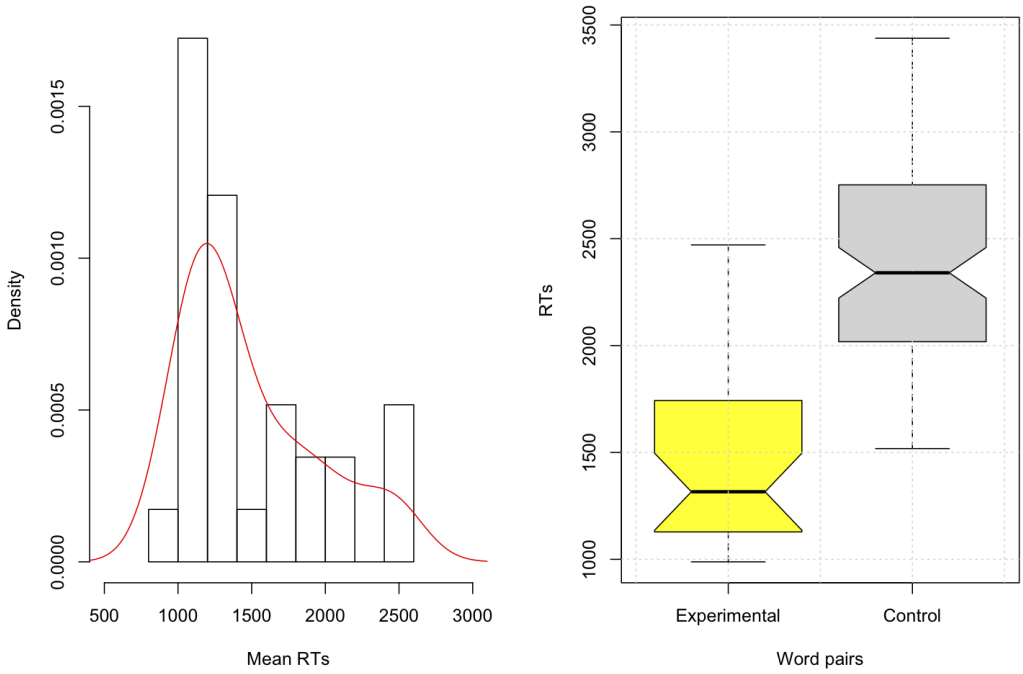


Figure 4.11: RTs and comparison of experimental vs. control word pairs: LEXG.

The comparison of RTs for LEXG and control pairs revealed a significant difference ($p < 0.001$; $W = 254.5$), attesting to the psychological salience of the word pairs selected by the AM (cf. Figure4.11, right panel). The (negative)

correlation between LEXG scores and RTs, however, is weak and only marginally significant (τ = -0.231, p = 0.081).

4.4.3.1.3 Log-likelihood. Word pairs selected by LL were recognized as plausible by participants in 88% of all cases. The mean of RTs was 1697 ms, the median 1514 ms and SD 654.7. Unlike FQ and LEXG, a long tail of RTs longer than 2000 ms was observed, with one value in the 3000-3500 ms range (cf. Figure 4.12, left panel). As for the other AMs, RTs are not normally distributed (p = 0.006, W = 0.8879).

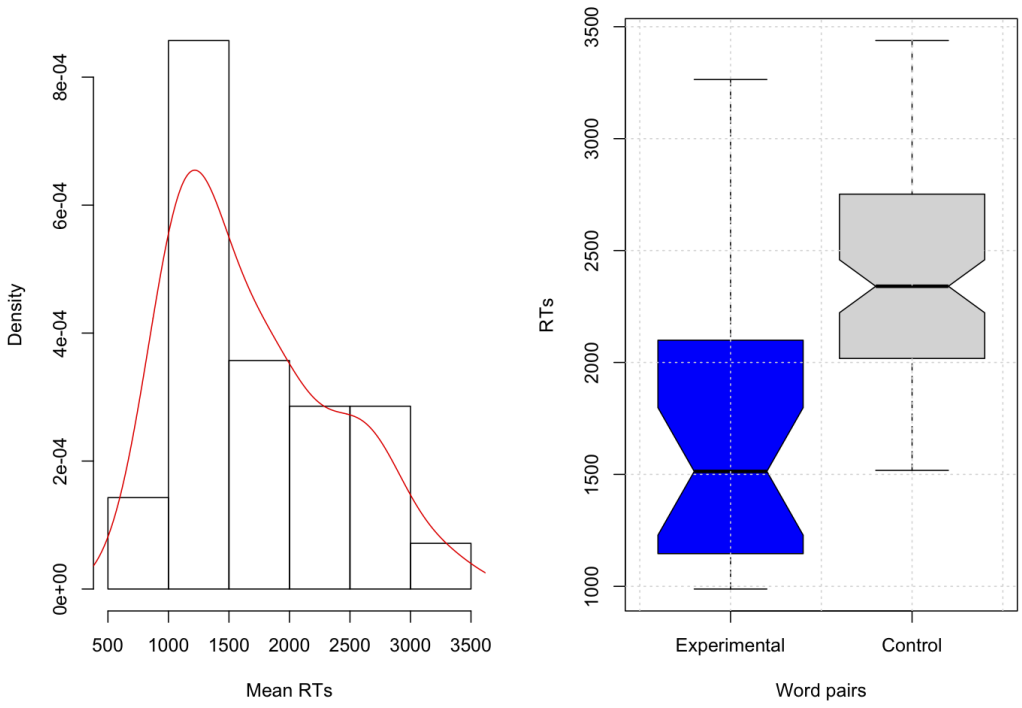


Figure 4.12: RTs and comparison of experimental vs. control word pairs: LL.

The trends highlighted for FQ and LEXG in the comparison of experimental and control pairs are also observed in the case of LL, whose RTs display significantly lower values than implausible pairs (p = 0.006, W = 0.8879). LL scores also display the strongest negative correlation with RTs observed so far (τ = -0.505, $p < 0.001$).

4.4.3.1.4 Mutual information. The accuracy with which MI word pairs were recognized as plausible was similar to that of LL (89%). The mean of RTs was 1720, the median 1576 and SD 533.6. As with LL, RTs reach 3000 ms (cf. Figure 4.13, left panel). In this case, data display a normal distribution ($p = 0.07$, $W = 0.9345$); however, since this was not the case for the RTs of all other AMs, parametric tests cannot be used for comparing them.

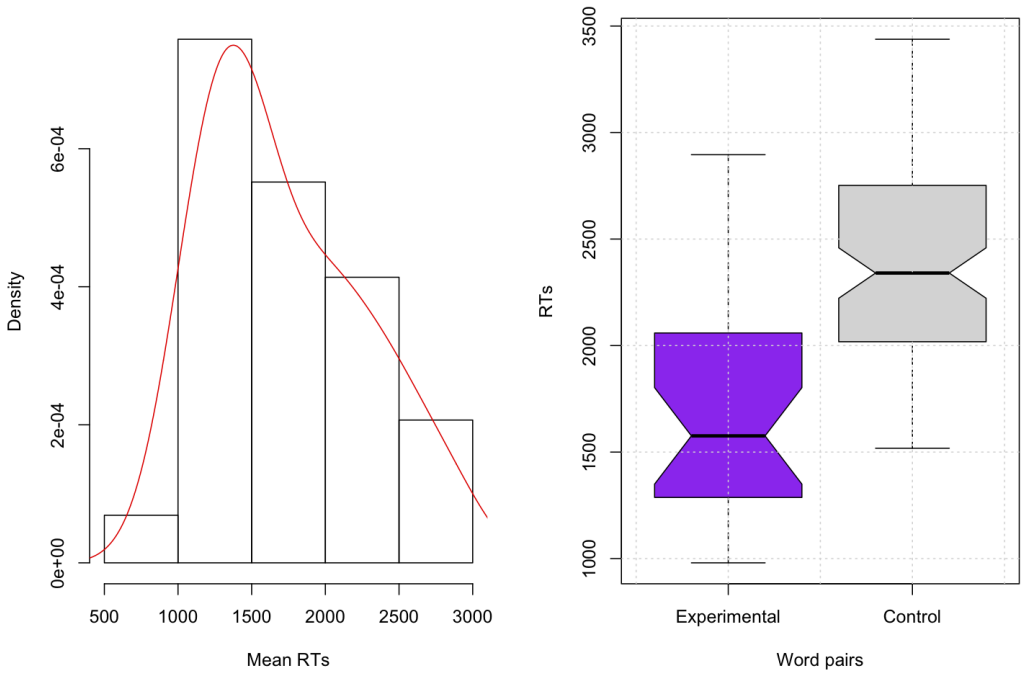


Figure 4.13: RTs and comparison of experimental vs. control word pairs: MI.

MI, like all other AMs, prompts significantly faster RTs when compared with control pairs ($p < 0.001$; $W = 489.5$; cf. Figure 4.13, right panel). However, no significant correlation was found between MI scores and RTs ($\tau = 0.037$, $p = 0.7781$).

4.4.3.2 Comparing the results: AMs and psycholinguistic data

The results presented in Section 4.4.3.1 showed that significant differences emerge between the RTs of word pairs selected by all the AMs and the control items, confirming that all the statistical measures extract psychologically salient pairs. The present Subsection takes a further step by comparing the results of the LDT

for the different AMs, with a view to assessing whether some are more likely to prompt faster recognition times than others, and hence extract word pairs which can be considered as *more* salient for NS of English.

AM	**Mean**	**Median**	**SD**	**Accuracy**	**Corr. values (AM score/RTs)**	
					τ	p
FQ	1416	1315	381.5	98%	-0.264	0.05
LEXG	1484	1316	460.2	95%	-0.231	0.08
LL	1697	1513	654.7	88%	-0.505	< 0.001
MI	1720	1576	533.6	89%	0.037	*ns*
ALL	1577	1388	525.3	93%	—	—
CONTROL	2392	2340	441.4	72%	—	—

Table 4.16: Descriptive statistics, accuracy scores and correlation values for the RTs (in *ms*) of the four AMs.

Table 4.16 provides descriptive statistics for each AM, as well as accuracy and correlation values (cf. Section 4.4.3.1). The distribution of RTs is presented in Figure 4.14: it would seem that AMs can be divided into two groups. On the one hand the RTs of FQ and LEXG display very similar distributions, with relatively low medians (compared to the other two AMs), and low SD values. It will also be remembered that the highest accuracy values were found for these AMs. In other words, it seems that FQ and LEXG prompt shorter RTs, and extract salient pairs more "reliably" (e.g. filtering out word pairs that are recognized as implausible by NS). On the other hand, the RTs of LL and MI also display similar distributions, but in this case median RT values are higher than those for FQ/LEXG, as are their SDs. It was hypothesized that these differences could be partly related to the higher number of cases in which LL and MI word pairs were recognized as implausible, which is reflected in lower accuracy scores, and which is likely to result in longer RTs. To test this hypothesis, a Mann-Whitney test was carried out, comparing the RTs of word pairs that the majority of participants (i.e. 6 or more) recognized as plausible (median = 1369 ms, SD = 480.4) with those that only a minority of them recognized as plausible (i.e. 5 or less; median = 2869, SD = 309.9): a significant difference was found, indicating that the former obtain shorter RTs than the latter ($p = 0.003$, $W = 3$). Of course, this test does not reveal whether the perceived implausibility of the word pairs is the *main factor* causing longer RTs. However, it seems fair to assume that since LL and MI extracted the highest number of implausible word sequences, this might have been one of the causes underlying *overall* higher mean/median and SD values for these AMs.

Returning to the comparison between the RTs associated with different AMs, the trends distinguishing FQ and LEXG on the one hand and LL and MI on the other were tested for statistical significance. Unlike in Sections 4.3.2.3-4.3.2.5, in

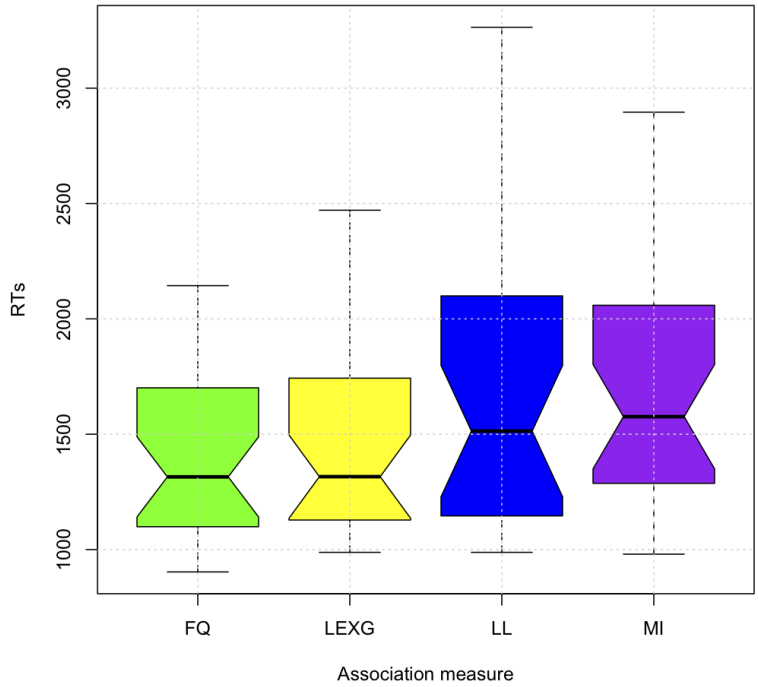


Figure 4.14: Reaction times associated with the word pairs selected by the four AMs.

this case it was not possible to use ANOVA as a significance test. The Bartlett statistics (cf. Gries (2009)) revealed that RT data violated one of the main assumptions on which this test is based, i.e. that the variances of the values associated with each factor (AMs) be homogeneous ($p = 0.03$; $K^2 = 8.59$). A non-parametric alternative to ANOVAs was therefore adopted, i.e. the Kruskal-Wallis rank sum test (cf. Gries (2009)).

Results indicate that differences between the RT values, considered as a function of AMs, only approached significance ($p = 0.1$; $H_3 = 5$). To further investigate whether pairwise comparisons between AMs' RT data return significant results, a series of Mann-Whitney tests was performed (cf. Table 4.17).

Only the difference between FQ and MI was found to be significant, while the difference between LEXG and MI approached significance. This finding lends support to the hypothesis that FQ (and to a minor extent LEXG, which has the highest correlation with FQ; cf. Section 3.4.2) prompts significantly shorter RTs than MI. This seems to contradict the results of Ellis and Simpson-Vlach (2009),

	FQ	LEXG	LL
LEXG	$p = ns$, $W = 407.5$	–	–
LL	$p = ns$, $W = 335.0$	$p = ns$, $W = 342.0$	–
MI	$\mathbf{p = 0.03}$, $W = 287.5$	$\mathbf{p = 0.06}$, $W = 298.5$	$p = ns$, $W = 371$

Table 4.17: Mann-Whitney test results for pairwise comparisons between the RTs of the four AMs.

who found that MI was the most solid predictor of RTs. It is true that due to the experimental design it was not possible to carry out the analysis of RTs using the same statistical techniques (multiple regression) adopted by the two authors (cf. Section 3.5.3). However, both correlation analysis and the results of pairwise comparisons between AMs indicate that, in this experiment, not only did FQ display a significant correlation with RTs, while MI did not, but it also prompted significantly shorter RTs than MI. As for LL, no comparison returned significant differences: its RTs occupy therefore a middle ground between the "two poles" represented by FQ/LEXG and MI.

4.4.4 Qualitative observations: the word pairs with the shortest and longest RTs

As was the case in the analysis of collocativity judgments (cf. Section 4.3.3.3), the insights derived from the quantitative analysis of RTs will be complemented by a more qualitative-oriented exploration. In what follows, the 10 word pairs with the shortest and longest RTs will be inspected: these should provide clues as to the nature of the word sequences which were associated with faster/slower processing on the part of the NSs.

The left panel of Table 4.18 displays the stimuli which prompted the shortest RTs. According to the lexicographic categorization that was also adopted to classify the collocation candidates with the highest/lowest collocativity ratings (cf. Section 4.3.3.3 and 4.2.2), these word pairs can be categorized as:

- free combinations/compositional phrases: "first year", "more information"; although not included in the dictionaries considered, "French novel", "third year" and "French politics" would (intuitively) seem to fall into this category as well;
- collocation-like sequences: "renewable energy", "serious illness";
- compounds: "black holes", "front line".

As can be observed, the majority of the word combinations associated with the shortest RTs are examples of the less "cohesive" types of word combinations,

with compounds and collocation-like sequences contributing 2 examples each. This would seem to contradict the idea that phraseologically-defined collocations are stored in the mental lexicon as units, while compositional phrases are "generated through the use of syntax and vocabulary", resulting in slower processing (Schmitt et al. 2004:128). These results are consistent with those of Schönefeld (2001:Ch. 6), who finds no significant difference in the RTs for "productive" and "fixed" phrases, roughly corresponding to the categories of compositional vs. collocation-like/compound sequences. However, in the same study, Schönefeld (2001) also finds that repeated exposure to a word as a stimulus leads to faster processing, irrespective of whether the word sequence in which it is included is a productive/fixed phrase. In our case, e.g., participants were presented with the word "French" 4 times as part of experimental pairs: this might explain the short RT for pairs like "French novel/politics". Further studies would therefore be required to (dis)confirm the findings described here.

Shortest RTs			**Longest RTs**		
Pair	**AM**	**RT (ms)**	**Pair**	**AM**	**RT (ms)**
French novel	FQ	903	design-based competition	LEXG	2471
renewable energy	MI	980	assessed individual	MI	2477
first year	FQ-LEXG-LL	988	cochlear implants	LL	2484
third year	FQ-LL	996	nearest halls	MI	2512
more information	FQ-LEXG-LL	1001	bioadhesive polymers	LL	2546
French politics	FQ	1025	sufficient sketchbooks	LL	2648
black holes	LL	1031	consultative committees	LL	2714
serious illness	LEXG	1039	articular cartilage	LL-MI	2744
front line	FQ	1048	reactive compatibilisers	MI	2896
key skills	LEXG	1054	automated DNA	LL	3264

Table 4.18: Word pairs with the shortest and longest RTs.

Turning to the analysis of the word pairs for which the *longest* RTs were observed (in the right panel of Table 4.18), results are hardly surprising: with the exception of "*cochlear implants*", the majority of the word pairs are examples of what we called "suspicious" collocation candidates (cf. 4.3.3.3), i.e.:

- word pairs occurring in pages of a single university: "design-based competition", "bioadhesive polymers", "consultative committees", "articular cartilage", "reactive compatibilisers";
- ill-formed word pairs: "assessed individual" (⇒ "assessed individual and group reports"), "nearest halls" (⇒ "nearest halls of residence"), "sufficient sketchbooks" (⇒ "sufficient. Sketchbooks"), "automated DNA" (⇒ "automated DNA sequencing").

It could be argued that these results are uninformative – they only show that NSs have longer RTs if a word pair is not well-formed/attested –, and that in a "traditional" psycholinguistic experiment, where stimuli are carefully controlled (cf. Section 2.3.4), such word sequences would probably not be included among the experimental stimuli. However, excluding them *a priori* would have run counter to the exploratory, corpus linguistics-oriented approach that is adopted in this thesis: in order to gain a better understanding of AMs and their performance, it was decided that their output should be manipulated as little as possible.

4.4.5 Psycholinguistic data and collocativity ratings

The last analysis that was carried out was aimed at assessing the degree of correlation between RTs and acceptability judgements obtained by expert informants (Section 4.3), namely *a)* whether the times required to process/recognize collocation candidates were predicted by experts' explicit intuition, and *b)* whether this effect was stronger for the word pairs selected by different AMs. For consistency, only the collocativity ratings provided by NS were taken into account.

As for question *a)* correlation analysis revealed that, overall, collocativity ratings were significantly correlated with RTs (cf. Table 4.19): the highest the ratings, the shortest was the time required to recognize collocation candidates. Converging evidence emerges therefore from the two tasks: expert intuitions *are* reliable predictors of mental processing.

AM	**Corr. values (RTs / Mean Rating)**	
	τ	p
OVERALL	-0.445	< 0.001
FQ	-0.438	< 0.001
LEXG	-0.443	< 0.001
LL	-0.448	< 0.001
MI	-0.57	< 0.001

Table 4.19: Correlation values for RTs and mean collocativity ratings (NSs only).

Turning to question *b)*, it would seem that the values of correlation strength are similar for all AMs, with FQ displaying the lowest correlation value, MI the highest, and LL and LEXG obtaining intermediate values. As can be observed in Figure 4.15, in the case of FQ (and to a lesser extent of LEXG and LL), the slightly weaker correlation seems to be mainly due to pairs that were judged as relatively non-salient by experts, often corresponding to free combinations such as "French politics" (see the discussion in 4.4.4).

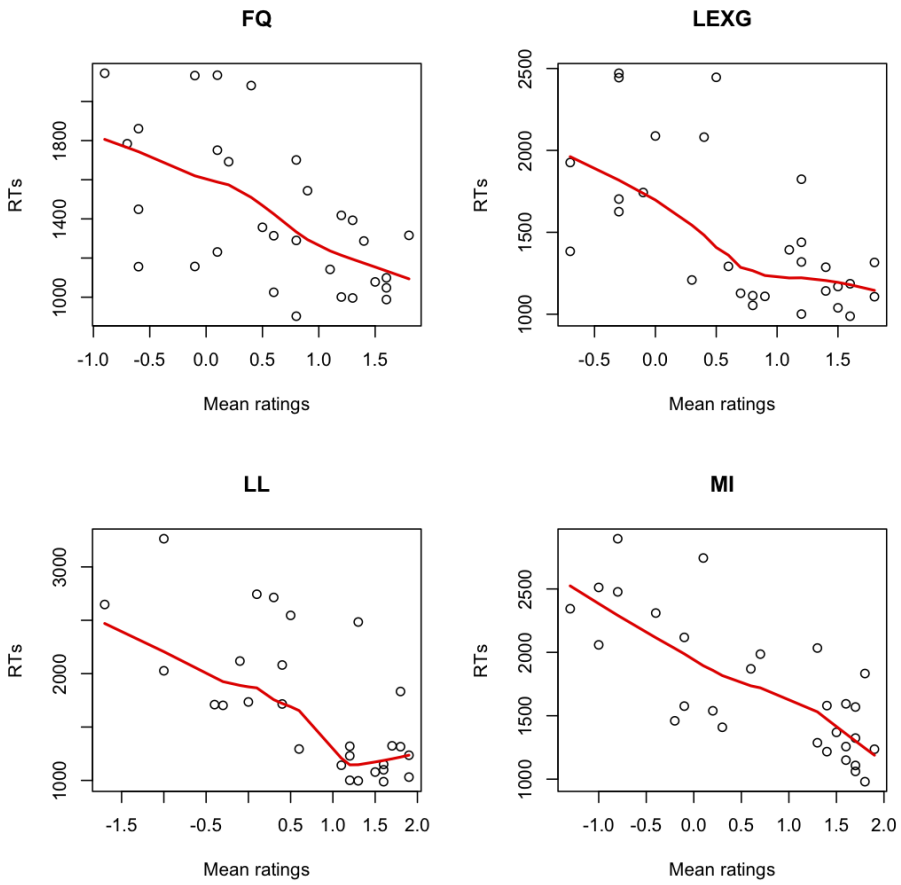


Figure 4.15: Correlation between RTs and acceptability judgements for the four AMs.

4.4.6 Interim summing up

A lexical decision task was set up involving 11 native speakers of English, who were presented with the 99 adjective-noun sequences of the evaluation set. Their RTs and the accuracy with which they responded to the stimuli were used as evidence to explore research question 2, i.e. whether "AMs predict the strength of association of word sequences in the minds of native speakers" and whether "a given AM predict it better than the others?" (3.2). Experimental evidence suggested that the collocation candidates extracted by all AMs were processed significantly faster than implausible (control) word pairs, thus witnessing to the idea that corpus evidence *is* reflected in competence/psycholinguistic processes. As for the ability of different AMs to predict word pair recognition times, the

scores of two AMs, i.e. FQ and LL, were found to be significantly correlated with RTs; for LEXG, correlation was only marginally significant, and for MI it was not significant. Moreover, a comparison of RTs across the pairs selected by different AMs demonstrated that sequences selected by FQ were processed significantly faster than those selected by MI.

The observation that FQ is the strongest predictor of RTs seems to contradict the results of Ellis and Simpson-Vlach (2009), and instead support those of Lapata et al. (1999). In Section 2.3.4.2, we speculated that the difference between the two previous studies was due to the different sampling strategies adopted – inspired by psycholinguistic concerns in the former case, and by corpus-linguistic concerns in the latter case. The results presented in this Section seem to (indirectly) support that speculation: when frequency is combined with a part-of-speech filter, it *is* a solid predictor of mental processes.

Finally, the analysis presented in Section 4.4.5 also makes it possible to provide a (tentative) answer to research question 3, i.e. "to what extent do corpus data, expert judgments, and experimental evidence provide converging evidence as to the phenomenon of collocation?". The answer seems to be that converging evidence did emerge from the three experiments: the expressions that experts found to be the most acceptable/most salient were also recognized faster and more accurately by subjects in the test. In turn, this experimental evidence is shown to be correlated with the corpus evidence extracted by the AMs.

4.4.7 Summing up

The present Chapter has presented the results of the three experiments which were set up for the evaluation of lexical association measures, involving lexicographic evidence, acceptability judgements obtained from expert informants, and psycholinguistic data reflecting mental processes of language comprehension/recognition. The analysis of data collected in each experiment was followed by a brief discussion highlighting the relevance of the analyses themselves for the purposes of the present work. The next Chapter summarizes the main insights gained from the research and concludes this thesis by presenting suggestions for future work.

Chapter 5

Conclusions and future work

5.1 General conclusions

The present thesis has covered the topic of lexical association measures (AMs) and their evaluation by triangulating corpus, lexicographic and experimental evidence. A specialized corpus of degree course descriptions, a well-defined genre within institutional academic English, was built adopting web-based, semi-automatic procedures. Focusing on the adjective-noun syntactic pattern, four AMs (namely frequency of co-occurrence, lexical gravity, log-likelihood and mutual information) were used to extract collocation candidates from this purpose-built corpus, adopting a stratified sampling technique. The 99 word pairs resulting from this procedure were evaluated by means of three evaluation tasks, i.e. an analysis of dictionary coverage, an acceptability judgement questionnaire, and a lexical decision task. Results have been presented and discussed at length in Chapter 4. The present Chapter draws some general conclusions, focusing on the main theoretical, methodological and applied/descriptive implications of these results. The thesis ends with suggestions for directions in which it the ressearch work presented here could progress.

In Section 2.3.3 it was argued that the performance-based approach to language investigation that characterizes the discipline of corpus linguistics – or, in the words of Leech (1992), its focus on language "as a product" – has led us to play down, or even overlook, the competence-related/psychological mechanisms underlying the linguistic phenomena observed. In the specific case of collocations and their identification in corpora through statistical measures, it is often *assumed* that collocativity scores, calculated on the basis of word frequencies in corpora, reflect psychological salience, though this assumption has been seldom tested empirically. The question was addressed explicitly in this thesis.

Relying on widely adopted methodologies in the psycholinguistics field, the word pairs extracted by different AMs and classified on the basis of lexicographic

coverage were evaluated against the salience of the word pairs as indicated by experts' judgements, and against the evidence of recognition/comprehension processes as indicated by the reaction times of native speakers in a lexical decision task. The results presented in Sections 4.3 and 4.4, suggest that a performance-based view of collocation reflects its competence-based counterpart. The collocativity scores assigned to word pairs by the three out of four AMs (with the exception of MI) were found to be able to predict both acceptability judgements and RTs. In turn, converging evidence was provided by the triangulation of corpus and experimental data: the expressions that experts found to be the most salient were also recognized faster and more accurately by subjects in the two tests.

Though not central to the main concerns of the investigation, an interesting result was obtained in the acceptability judgement task concerning differences in the way native and non-native speakers (NSs and NNSs) of English evaluated the word pairs in the questionnaire. The latter were found to provide less extreme collocativity ratings than the former (i.e. ratings tending to 0), which in previous research was hypothesized to be an indication of non-native speakers' collocational knowledge being substantially less developed than that of natives (Granger 1998; Siyanova and Schmitt 2008). An alternative explanation was tentatively suggested in Section 4.3.3.4, based on the observation that *a)* native and non-native speakers' judgements displayed almost identical correlation values to scores assigned by the different AMs, and *b)* similar levels of consensus, measured by an inter-rater agreement statistic, were obtained for the two groups. In particular, it was suggested that the less extreme ratings provided by NNSs might be the result of strategies of "risk aversion" or, in other words, that non-native informants tend not to provide very positive or very negative ratings because they are less confident than native informants, even though the intuitions of both groups on collocativity substantially match. The issue could not be settled here, and several variables might have influenced the results (e.g. the imbalances in the number of participants belonging to the two groups), but this finding certainly warrants further investigation.

On the methodological side, several insights were gained concerning the performance of AMs in automatically identifying collocations in corpora. In the lexicographic task and the acceptability judgement task, quantitative analyses showed that no AM clearly outperformed the others in extracting "salient collocations", a result that is consistent with previous research (e.g. Evert and Krenn 2001). A more qualitative-oriented analysis, however, suggested that *a)* differences emerge in terms of the *types* of word combinations that different AMs give prominence to, and *b)* that these word combinations tend to prompt diverging judgements on their salience: FQ tended to extract free, compositional word combinations, while MI targeted restricted-collocation-like sequences and

compounds. LEXG and LL were found to occupy a middle ground between the other two measures. A last point worth noting concerns the stratified sampling method adopted in this thesis, which resulted in word pairs being extracted from the "n-best" set for each measure, as well as from the top, middle and bottom frequency ranges. On the basis of expert judgements, one can conclude that the best candidates selected by FQ and LEXG are indeed more likely to be salient since they consistently received (significantly) higher ratings. The same does not hold for LL and especially MI, though. The best LL pairs are not significantly better than those in the frequency ranges, while in the case of MI the highest ratings are obtained by pairs in the high frequency range, even though, as is well known, the statistic tends to play up "surprising", low frequency pairs.

These results have applied implications. When deciding which AM to use for a collocation extraction task, it is crucial that the purpose of the extraction be considered. If collocations are extracted, e.g., for inclusion in a dictionary or termbase following manual inspection, MI can provide unexpected, very salient phrases (particularly if one adopts a frequency threshold). However, these would have to be sifted out of a large number of ill-formed phrases or casual sequences. If, on the other hand, the purpose is to provide a fully automatic list of collocations (e.g. for machine translation or other NLP tools that aim to facilitate writing or translation into a foreign language), then one would be better off relying on FQ or LEXG, selecting for analysis the best n word pairs identified by the measures.

While this thesis has a clear methodological focus, its starting point was in fact an applied one, namely the identification of methods for extracting phraseology from ESP corpora, in particular in the field of institutional academic English. As a result of the work conducted, a pipeline has been set up for replicating the corpus construction phase and an understanding of typical phraseology in this area, and of the most appropriate ways of extracting it from corpora, has been developed. The collocation candidates extracted show that the phraseology of this genre is characterized by a mix of disciplinary terms (such as "cochlear implants"), "common core" expressions ("open days") and more or less interesting general language sequences ("front line", "beautiful city"). By tweaking the parameters used for collocation extraction it might be possible to focus specifically on each one of these sets of phrases and proceed to an investigation of their usage in context. The subsequent step would be to extend the corpus to include the *lingua franca* variety of the same ESP, as well as other genres within this ESP. In this way we could start to understand the main phraseological differences between native and non-native / translated text production in this field, as a first step towards making sure that the latter is (at least) as communicatively effective as the former.

Other directions in which this work could be extended include tapping the

competence of NNSs and learners more systematically through psycholinguistic experiments matching the questionnaire survey (that included NNSs' judgements). This perspective could shed light on differences between NSs and NNSs both in terms of their use of collocations (as displayed by corpora) and of their intuition about, and implicit knowledge of, collocations (as inferrable from questionnaires and psycholinguistic experiments). Finally, the comparative evaluation of the different AMs could be extended to include a practical NLP task. Automatic synonymy detection would seem to be especially apt for this purpose, given the role collocations play within distributional semantics views of synonymy (Curran 2004). By triangulating data from as many approaches as possible – product- and process-oriented, theoretical, descriptive and applied – it might be possible to finally start to disentangle some of the complexities underlying the notion of collocation.

References

Afros, E. and C. F. Schryer (2009). The genre of syllabus in higher education. *Journal of English for Academic Purposes 8*(3), 224–233.

Altbach, P. G. and J. Knight (2007). The internationalization of higher education: Motivations and realities. *Journal of Studies in International Education 11*(3-4), 290–305.

Artstein, R. and M. Poesio (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics 34*(4), 555–596.

Baayen, H. R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London and New York: Continuum.

Baroni, M. and S. Bernardini (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, Lisbon, Portugal, pp. 1313–1316. ELDA.

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta (2009). The Wacky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation 43*(3), 209–226.

Bartsch, S. (2004). *Structural and functional properties of collocations in English*. Tübingen: Gunter Narr.

Benson, M. (1985). Collocations and idioms. In R. Ilson (Ed.), *Dictionaries, lexicography and language learning*, pp. 61–68. Oxford: British Council/Pergamon Press.

Benson, M. (1989). The structure of the collocational dictionary. *International Journal of Lexicography 2*(1), 1–14.

Bernardini, S. (2007). *Collocations in translated text. A corpus-based study.* Ph. D. thesis, Middlesex University.

Bernardini, S., M. Baroni, and S. Evert (2006). A WaCky introduction. In M. Baroni and S. Bernardini (Eds.), *WaCky! Working Papers on the Web as Corpus*, pp. 9–40. Bologna: GEDIT.

Bernardini, S. and A. Ferraresi (forthcoming). Old needs, new solutions – Comparable corpora for language professionals. In S. Sharoff, R. Rapp, P. Zweigenbaum, and P. Fung (Eds.), *Building and Using Comparable Corpora*. Dordrecht: Springer.

Bernardini, S., A. Ferraresi, and F. Gaspari (2010). Institutional academic English in the European context: A web-as-corpus approach to comparing native and non-native language. In A. Linde López and R. Crespo Jiménez (Eds.), *Professional English in the European context: The EHEA challenge*, pp. 27–53. Bern: Peter Lang.

Berry-Rogghe, G. L. M. (1973). The computation of collocations and their relevance to lexical studies. In A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith (Eds.), *The Computer and Literary Studies*, pp. 103–112. Edinburgh: Edinburgh University Press.

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings.* Harlow: Longman.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan (1999). *Longman grammar of spoken and written English.* London: Longman.

Burnard, L. (1995). *The British National Corpus user's reference guide.* Oxford: Oxford University Computing Services.

Caiazzo, L. (2010). The 'promotional' English(es) of university websites. In R. Cagliero and J. Jenkins (Eds.), *Discourses, communities, and global Englishes*, pp. 43–60. Bern: Peter Lang.

Carter, R. (1998). *Vocabulary. Applied Linguistic Perspectives* (2nd ed.). London and New York: Routledge.

Choueka, Y. (1988). Looking for needles in a haystack. In *Proceedings of RIAO '88*, pp. 609–623.

Church, K. W. and P. Hanks (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics 16*(1), 22–29.

Clark, H. H. (1970). Word associations and linguistic theory. In J. Lyons (Ed.), *New horizons in linguistics*, pp. 271–286. Harmondsworth: Penguin.

Cohen, J. D., B. MacWhinney, M. Flatt, and P. J. (2006). Psyscope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers 25*(2), 103–129.

Connor, U. and T. A. Upton (2004). The genre of grant proposals: A corpus linguistic analysis. In U. Connor and T. A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics*, pp. 235–256. Amsterdam and Philadelphia: John Benjamins.

Cowie, A. P. (1978). The place of illustrative material and collocations in the design of a learner's dictionary. In P. Strevens (Ed.), *In Honour of A.S. Hornby*, pp. 127–139. Oxford: Oxford University Press.

Cowie, A. P. (1988). Stable and creative aspects of vocabulary use. In R. Carter and M. McCarthy (Eds.), *Vocabulary and language teaching*, pp. 126–139. Harlow: Longman.

Cowie, A. P. (1998a). Introduction. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications*, pp. 1–20. Oxford: Oxford University Press.

Cowie, A. P. (Ed.) (1998b). *Phraseology: Theory, analysis, and applications*, Oxford. Oxford University Press.

Croft, W. and A. D. Cruse (2004). *Cognitive linguistics*. Cambridge: Cambridge University Press.

Curran, J. R. (2004). *From Distributional to Semantic Similarity*. Ph. D. thesis, University of Edinburgh, Edinburgh.

Cermák, F. (2006). Collocations, collocability and dictionary. In E. Corino, C. Marello, and C. Onesti (Eds.), *Proceedings XII EURALEX International Congress*, Alessandria, pp. 929–937. Edizioni dell'Orso.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : Statistiques lexicales et filtres linguistiques*. Ph. D. thesis, Université Paris 7.

Danielsson, P. (2003). Automatic extraction of meaningful units from corpora. A a corpus-driven approach using the word *stroke*. *International Journal of Corpus Linguistics 8*(1), 109–127.

Daudaravičius, V. and R. Marcinkevičiene (2004). Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics 9*(2), 321–348.

Depraetere, H., J. Van den Bogaert, and J. Van de Walle (2011, May). Bologna translation service: Online translation of course syllabi and study programmes in English. In M. L. Forcada, H. Depraetere, and V. Vandeghinste (Eds.), *Proceedings of the 15th conference of the European Association for Machine Translation*, Leuven, Belgium, pp. 29–34.

Drew, P. and J. Heritage (Eds.) (1992). *Talk at work: Interaction in institutional settings*, Cambridge. Cambridge University Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics 19*(1), 61–74.

Durrant, P. and A. Doherty (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory 6*(2), 125–155.

Ellis, N., R. Simpson-Vlach, and C. Maynard (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly 42*(3), 375–396.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition 24*(2), 143–188.

Ellis, N. C. and R. Simpson-Vlach (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory 5*(1), 61–78.

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. Ph. D. thesis, Universität Stuttgart, Stuttgart.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics*, Volume 2, pp. 1212–1248. Berlin, New York: Mouton de Gruyter.

Evert, S. and B. Krenn (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th annual meeting of ACL*, Stroudsburg, PA (USA), pp. 188–195. Association for Computational Linguistics.

Fairclough, N. (1993). Critical discourse analysis and the marketization of public discourse: The universities. *Discourse & Society 4*(2), 133–168.

Fairon, C., H. Naets, A. Kilgarriff, and G.-M. de Schryver (Eds.) (2007). *Proceedings of the 3rd Web-as-Corpus workshop, incorporating Cleaneval*, Louvain-la-Neuve. Presses universitaires de Louvain.

Fantinuoli, C. (2006). Specialized corpora from the web and term extraction for simultaneous interpreters. In M. Baroni and S. Bernardini (Eds.), *WaCky! Working Papers on the Web as Corpus*, pp. 173–190. Bologna: GEDIT.

Fazly, A., S. Stevenson, and R. North (2007). Automatically learning semantic knowledge about multiword predicates. *Language Resources and Evaluation 41*(1), 61–89.

Ferraresi, A. and S. T. Gries (2011). Type and (?) token frequencies in measures of collocational strength: Lexical gravity vs. a few classics. In *Paper presented at Corpus Linguistics 2011*, Birmingham (UK). University of Birmingham.

Fillmore, C., P. Kay, and C. O'Connor (1988). Regularity and idiomaticity in grammatical constructions: The case of "let alone". *Language 64*, 501–538.

Firth, J. R. (1956/1968a). Descriptive linguistics and the study of English. In F. R. Palmer (Ed.), *Selected papers of J.R.Firth 1952-1959*, pp. 96–113. Harlow: Longman.

Firth, J. R. (1956/1968b). A synopsis of linguistic theory 1930-55. In F. R. Palmer (Ed.), *Selected papers of J. R. Firth, 1952-59*, pp. 1–32. Harlow: Longman.

Fitzpatrick, T. (2007). Word association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics 3*(17), 319–331.

Fletcher, W. (2004). Making the web more useful as a source for linguistic corpora. In U. Connor and T. Upton (Eds.), *Corpus Linguistics in North America 2002*, pp. 191–205. Amsterdam: Rodopi.

Fox, G. (1987). The case for examples. In J. M. Sinclair (Ed.), *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English Language Dictionary*, pp. 137–149. London and Glasgow: Collins.

Gatto, M. (2009). *From Body to Web. An Introduction to the Web as Corpus*. Bari: Laterza.

Gesuato, S. (2011). Course descriptions: Communicative practices of an institutional genre. In S. Sarangi, V. Polese, and G. Caliendo (Eds.), *Genre(s) on the Move: Hybridization and Discourse Change in Specialized Communication*, pp. 221–241. Napoli: Edizioni Scientifiche Italiane.

Gilquin, G. (2008). What you think ain't what you get: Highly polysemous verbs in mind and language. In J.-R. Lapaire, G. Desagulier, and J.-B. Guignard (Eds.), *Du fait grammatical au fait cognitif. From gram to mind: Grammar as cognition*, Volume 2, pp. 235–255. Pessac: Presses Universitaires de Bordeaux.

Gilquin, G. and S. T. Gries (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory 5*(1), 1–26.

Gitsaki, C. (1996). *The development of ESL collocational knowledge*. Unpublished doctoral dissertation, University of Queensland, Brisbane.

Gläser, R. (1988). The grading of idiomaticity as a presupposition for a taxonomy of idioms. In W. Hüllen and R. Schulze (Eds.), *Understanding the Lexicon: Meaning, Sense and World Knowledge in Lexical Semantics*, pp. 264–279. Tübingen: Max Niemeyer.

Godfrey, J. J. and E. Holliman (1997). *Switchboard-1. Release 2*. Philadelphia: Linguistic Data Consortium.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications*, pp. 145–160. Oxford: Oxford University Press.

Granger, S. and M. Paquot (2008). Disentangling the phraseological web. In S. Granger and F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective*, pp. 27–49. Amsterdam and Philadelphia: John Benjamins.

Gries, S. T. (2008). Phraseology and linguistic theory. a brief survey. In S. Granger and F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective*, pp. 3–26. Amsterdam and Philadelphia: John Benjamins.

Gries, S. T. (2009). *Statistics for linguistics with R. A practical introduction*. Berlin and New York: Mouton de Gruyter.

Gries, S. T. (2010a). Bigrams in registers, domains, and varieties: A bigram gravity approach to the homogeneity of corpora bigram gravity approach to the homogeneity of corpora. In *Proceedings of Corpus Linguistics 2009*, University of Liverpool.

Gries, S. T. (2010b). Useful statistics for corpus linguistics. In A. Sánchez and M. Almela (Eds.), *A mosaic of corpus linguistics: selected approaches*, pp. 269–291. Frankfurt am Main: Peter Lang.

Gries, S. T. and J. Mukherjee (2010). Lexical gravity across varieties of English: an ice-based study of n-grams in asian Englishes. *International Journal of Corpus Linguistics 15*(4), 520–548.

Halliday, M. A. K. (1961). Categories of the theory of grammar. *Word 17*(3), 241–292.

Halliday, M. A. K. and R. Hasan (1976). *Cohesion in English*. London: Longman.

Handl, S. (2008). Essential collocations for learners of English: The role of collocational direction and weight. In F. Meunier and S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching*, pp. 43–66. Amsterdam and Philadelphia: John Benjamins.

Hausmann, F. J. (1989). Le dictionnaire de collocations. In H. E. W. Hausmann and L. Zgusta (Eds.), *Wörterbücher, dictionaries, dictionnaires. Ein internationales Handbuch zur Lexikographie*, pp. 1010–1019. Berlin: Mouton de Gruyter.

Hausmann, F. J. (1997). Tout est idiomatique dans les langues. In M. Martins-Baltar (Ed.), *La locution entre langue et usages*, pp. 277–290. Fontenay-Saint Cloud: ENS éditions.

Hausmann, F. J. (1999). Le dictionnaire de collocations: Critères de son organisation. In N. Greiner, J. Kornelius, and G. Rovere (Eds.), *Texte und Kontexte in Sprachen und Kulturen. Festschrift für Jörn Albrecht*, pp. 121–139. Trier: WVT.

Hausmann, F. J. and P. Blumenthal (2006). Présentation: Collocations, corpus, dictionnaires. In P. Blumenthal and F. J. Hausmann (Eds.), *Collocations, corpus, dictionnaires*, Number 149 in Langue française, pp. 3–13. Paris: Armand Colin.

Hoey, M. (2005). *Lexical priming*. London and New York: Routledge.

Howarth, P. A. (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making*. Tübingen: Max Niemeyer.

Hunston, S. (2001). Colligation, lexis, pattern and text. In M. Scott and G. Thompson (Eds.), *Patterns of text. In honour of Michael Hoey*, pp. 13–34. Amsterdam and Philadelphia: John Benjamins.

Hyland, K. (2011). Projecting an academic identity in some reflective genres. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos 21*, 9–30.

Jenkins, J. (2007). *English as a Lingua Franca: Attitude and identity*. Oxford: The University of Michigan Press.

Johansson, S. (1993). Data, description, discourse. In M. Hoey (Ed.), *Sweetly oblivious: Some aspects of adverb-adjective combinations in present-day English*, pp. 39–49. London: Collins.

Johansson, S., E. Atwell, R. Garside, and G. Leech (1986). *The tagged LOB corpus - Users' manual*. Bergen: Norwegian Computing Centre for the Humanities.

Jones, S. and J. M. Sinclair (1974/1996). English lexical collocations. In J. A. Foley (Ed.), *Sinclair on lexis and lexicography*, pp. 21–54. Singapore: UniPress.

Kachru, B. (1985). Standards, codification and sociolinguistic realism: The english language in the outer circle. In R. Quirk and H. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures*, pp. 11–30. Cambridge: Cambridge University Press.

Kennedy, G. (1992). Preferred ways of putting things with implications for language teaching. In J. Svartvik (Ed.), *Directions in corpus linguistics*, pp. 335–373. Berlin: Mouton de Gruyter.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics 6*(1), 97–133.

Kilgarriff, A. and D. Tugwell (2002). Sketching words. In M.-H. Corréard (Ed.), *Lexicography and natural language processing: A festschrift in honour of B. T. S. Atkins*, pp. 125–137. Grenoble: EURALEX.

Kjellmer, G. (1991). A mint of phrases. In K. Aijmer and B. Altenberg (Eds.), *English corpus linguistics. Studies in honour of Jan Svartvik*, pp. 111–127. London and New York: Longman.

Krenn, B. and S. Evert (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, Toulouse, France, pp. 39–46.

Lapata, M., S. McDonald, and F. Keller (1999). Determinants of adjective-noun plausibility. In *Proceedings of the 9th conference of the European chapter of the Association for Computational Linguistics*, Morristown (NJ), pp. 30–36. Association for Computational Linguistics.

LDOCE (2003). *Longman Dictionary of Contemporary English*. Harlow: Longman.

Lea, D. and M. Runcie (2002, August 13-17). Blunt instruments and fine distinctions. A collocations dictionary for students of English. In A. Braasch and C. Povlsen (Eds.), *Proceedings of the Tenth EURALEX International Conference*, Copenhagen (DK), pp. 819–829. EURALEX.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics*, pp. 105–126. Berlin: Mouton de Gruyter.

Lewis, M. (1993). *The lexical approach*. Hove: Language Teaching Publications.

Mason, O. (1999). Parameters of collocation: The word in the centre of gravity. In J. M. Kirk (Ed.), *Corpora galore: Analyses and techniques in describing English*, pp. 267–280. Amsterdam: Rodopi.

Mauranen, A. (2003). Academic English as Lingua Franca - A corpus approach. *TESOL Quarterly 37*, 513–527.

Mautner, G. (2005). For-profit discourse in the nonprofit and public sectors. In G. Erreygers and G. Jacobs (Eds.), *Language, communication and the economy*, pp. 25–44. Amsterdam: John Benjamins.

McEnery, T., R. Xiao, and Y. Tono (2006). *Corpus-based language studies. An advanced resource book*. London and New York: Routledge.

McGee, I. (2009). Adjective-noun collocations in elicited and corpus data: Similarities, differences, and the whys and wherefores. *Corpus Linguistics and Linguistic Theory 5*(1), 79–103.

McKeown, K. R. and D. R. Radev (2000). Collocations. In R. Dale, H. Moisl, and H. Somers (Eds.), *A handbook of Natural Language Processing*, pp. 507–523. New York: Marcel Dekker.

Mel'čuk, I. (1988). Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography 1*(3), 165–188.

Mel'čuk, I. (1998). Collocations and lexical functions. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications*, pp. 23–53. Oxford: Oxford University Press.

Mollin, S. (2009). Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory 5*(2), 175–200.

Moon, R. (1998a). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Oxford University Press.

Moon, R. (1998b). Frequencies and forms of phrasal lexemes in English. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications*, pp. 79–100. Oxford: Oxford University Press.

Moon, R. (2008). Dictionaries and collocation. In S. Granger and F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective*, pp. 313–336. Amsterdam and Philadelphia: John Benjamins.

Murphy, B. (2007). *A Study of Notions of Participation and Discourse in Argument Structure Realisation*. Ph. D. thesis, Department of Computer Science, Trinity College Dublin.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam and Philadelphia: John Benjamins.

Nordquist, D. (2009). Investigating elicited data from a usage-based perspective. *Corpus Linguistics and Linguistic Theory 5*(1), 105–130.

O'Dell, F. and M. McCarthy (2008). *English collocations in use: Advanced*. Cambridge: Cambridge University Press.

Partington, A. (1998). *Patterns and Meanings*. Amsterdam: John Benjamins.

Pawley, A. and F. H. Syder (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards and R. Schmidt (Eds.), *Language and communication*, pp. 191–225. London: Longman.

Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. In *Proceedings of LREC 2002*, Las Palmas, Spain, pp. 1530–1536.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation 44*(1), 137–158.

Pluymaekers, M., M. Ernestus, and H. R. Baayen (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica 62*(2), 146–159.

Pym, A. (2008). On Toury's laws of how translators translate. In A. Pym, M. Shlesinger, and D. Simeoni (Eds.), *Beyond descriptive translation studies. Investigations in honor of Gideon Toury*, pp. 311–328. Amsterdam: Benjamins.

Renouf, A. and J. M. Sinclair (1991). Collocational frameworks in English. In K. Aijmer and B. Altenberg (Eds.), *English corpus linguistics. Studies in honour of Jan Svartvik*, pp. 128–143. London: Longman.

Römer, U. (2010). Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction 3*(1), 95–119.

Sag, I., T. Baldwin, F. Bond, A. Copestake, and D. Flickinger (2002). Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, Volume 2276 of *Lecture Notes in Computer Science*, pp. 189–206. Berlin, Heidelberg: Springer.

Schmitt, N. (1998). Quantifying word association responses: What is nativelike? *System 26*, 389–401.

Schmitt, N., S. Grandage, and S. Adolphs (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic Sequences*, pp. 127–151. Amsterdam: John Benjamins.

Schönefeld, D. (2001). *Where lexicon and syntax meet.* Berlin: Mouton de Gruyter.

Seidlhofer, B. (2001). Closing a conceptual gap: The case for a description of English as a Lingua Franca. *International Journal of Applied Linguistics 11*, 133–158.

Seretan, V. (2008). *Collocation extraction based on syntactic parsing.* Ph. D. thesis, University of Geneva.

Shaoul, C. and C. Westbury (2011). Formulaic sequences, do they exist and do they matter? *The Mental Lexicon 6*(1), 171–196.

Siepmann, D. (2005). Collocation, colligation and encoding dictionaries. Part I: Lexicological aspects. *International Journal of Lexicography 18*(4), 409–443.

Simpson-Vlach, R. C. and S. Leicher (2006). *The MICASE Handbook: A resource for users of the Michigan Corpus of Academic Spoken English.* Ann Arbor: The University of Michigan Press.

Sinclair, J. M. (1966). Beginning the study of lexis. In J. A. Foley (Ed.), *Sinclair on lexis and lexicography*, pp. 1–20. Singapore: UniPress.

Sinclair, J. M. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, J. M. (1996). The search for units of meaning. *Textus 9*(1), 75–106.

Sinclair, J. M. (1998). The lexical item. In E. Weigand (Ed.), *Contrastive lexical semantics*, pp. 1–24. Amsterdam: John Benjamins.

Sinclair, J. M. (2004). Corpus and text – Basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice*, pp. 1–16. Oxford: Oxbow Books.

Sinclair, J. M. and J. Ball (1996). EAGLES preliminary recommendations on text typology. Online: www.ilc.cnr.it/EAGLES/texttyp/texttyp.html [consulted 22.08.07].

Sinclair, J. M., S. Jones, R. Daley, and R. Krishnamurthy (2004/1970). *English collocation studies: The Osti report*. London: Continuum.

Siyanova, A. and N. Schmitt (2008). L2 learner production and processing of collocation: A multi-study perspective. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes 64*(3), 429–458.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational linguistics 19*(1), 143–177.

Spooren, W. and L. Degand (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory 6*(2), 241–266.

Stefanowitsch, A. and S. T. Gries (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics 8*(2), 209–243.

Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford: Blackwell.

Stubbs, M. (2001). *Words and phrases. Corpus studies in lexical semantics*. Oxford: Blackwell.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics 7*(2), 215–244.

Stubbs, M. (2009). Memorial article: John Sinclair (1933-2007). The search for units of meaning: Sinclair on empirical semanticshe search for units of meaning: Sinclair on empirical semantics. *Applied Linguistics 30*(1), 115–137.

Svensson, M. H. (2008). A very complex criterion of fixedness: Non-compositionality. In S. Granger and F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective*, pp. 81–93. Amsterdam and Philadelphia: John Benjamins.

Swales, J. (1990). *Genre analysis. English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M. (1987, March 19-21). Approaching the concept of discourse community. Paper presented at the annual meeting of the conference on College composition and communication.

Swales, J. M. (2004). *Research genres. Explorations and applications*. Cambridge: Cambridge University Press.

Tognini-Bonelli, E. (2004). *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John Benjamins.

Tutin, A. and F. Grossmann (2002). Collocations régulières et irrégulières: Esquisse de typologie du phénomène collocatif. *Revue française de linguistique appliquée 7*, 7–25.

Underwood, G., N. Schmitt, and A. Galpin (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic Sequences*, pp. 153–172. Amsterdam: John Benjamins.

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory 4*(2), 253–290.

Willis, D. (2001). *The lexical syllabus: A new approach to language teaching*. London: Collins.

Wolter, B. and H. Gyllstad (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics 32*(4), 430–449.

Wong, W., W. Liu, and M. Bennamoun (2011). Constructing specialised corpora through analysing domain representativeness of websites. *Language Resources and Evaluation 45*(2), 209–241.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wulff, S. (2008). *Rethinking idiomaticity: A usage-based approach*. London and New York: Continuum.

Xiao, R. and T. McEnery (2006). Collocation, semantic prosody and near synonymy: A cross-linguistic perspective. *Applied Linguistics 27*(1), 103–129.

Appendix A

Acceptability judgement questionnaire

QUESTIONNAIRE

INSTRUCTION SHEET

Adriano Ferraresi,
PhD student

University of Naples "Federico II" / University of Bologna
ITALY

Contacts:
E-mail: a.ferraresi@unina.it
Phone: +39 3332063104

Dear ICAMEr,

The present questionnaire is about **English multiword expressions**, loosely defined as **the kind of phrases whose members are characterized by a degree of lexical association**.

In the next two pages you will find **99 *adjective* + *noun* sequences** extracted from a genre-specific corpus of **undergraduate course descriptions**. I ask you to consider them in turn and assign each one **a score from 1 to 5** according to your perception of the degree of lexical association between its members, i.e. according to how strongly, in your opinion, the two words are attracted to each other and to what degree they form an **intuitively salient, interesting phrase**.

The scores 1 and 2 correspond to low degrees of lexical association ("no or very weak"; "weak"), and 4 and 5 to high degrees of lexical association ("strong"; "very strong"); if you are uncertain about the status of a sequence, assign it a "medium" value of 3.

In the "Comments" field you can **motivate your decision for any of the sequences**. You will find that **six sequences are highlighted**: it would be extremely valuable for this study if you could **comment at least on them**.

Completing the questionnaire should take no longer than **10/15 minutes**. I am interested in **your first impression**, so please do not spend too much time pondering your judgment.

Thank you very much for your help!

Adriano Ferraresi

Lexical association. 1: no, or very weak; 2: weak; 3: uncertain; 4: strong; 5: very strong

No.	Word pair	Lexical association	Comments?
1	French romanticism		
2	historic buildings		
3	practical work		
4	nucleic acids		
5	foundation-year entry		
6	active staff		
7	distinguished scholars		
8	reflexive individuals		
9	manual dexterity		
10	consultative committees		
11	linear algebra		
12	practical skills		
13	stand-up comedy		
14	worth two-thirds		
15	more information		
16	**naked eye**		
17	further information		
18	work-related learning		
19	finished product		
20	recommended GCSEs		
21	differential equations		
22	widest ranges		
23	serious illness		
24	automatic progression		
25	smooth transition		
26	wide range		
27	thermal conversion		
28	partial exemption		
29	proficient enough		
30	whole spectrum		
31	French law		
32	subatomic particles		
33	bioadhesive polymers		
34	black holes		
35	liquid bio-fuels		
36	web-based systems		
37	renewable energy		
38	academic training		
39	**final year**		
40	unequalled concentration		
41	acceptable subject		
42	simple notes		
43	transferable skills		
44	nearest halls		
45	assessed individual		
46	optional modules		
47	**cochlear implants**		
48	key skills		
49	articular cartilage		
50	design-based competition		
51	dynamic region		
52	initial concept		

Lexical association. 1: no, or very weak; 2: weak; 3: uncertain; 4: strong; 5: very strong

53	departmental website		
54	**open days**		
55	naval architecture		
56	reactive compatibilisers		
57	front line		
58	typical offer		
59	international students		
60	connecting UEL		
61	domestic animals		
62	binding agreement		
63	actual amount		
64	fresh insights		
65	fast pyrolysis		
66	volcanic eruptions		
67	white man		
68	work-based experience		
69	consistent representation		
70	weekly timetable		
71	architectural practices		
72	francophone country		
73	dramatic text		
74	strict deadlines		
75	second year		
76	video games		
77	cystic fibrosis		
78	coral reefs		
79	certified proof		
80	sufficient sketchbooks		
81	overriding goal		
82	first year		
83	subject area		
84	French politics		
85	automated DNA		
86	**beautiful city**		
87	manufactured goods		
88	additional tests		
89	French novel		
90	**rigid deadlines**		
91	one-day symposia		
92	responsible investment		
93	financial support		
94	real-life scenarios		
95	gross salary		
96	former graduates		
97	francophone world		
98	third year		
99	premier venues		

Lastly, a question about yourself - feel free not to answer if you prefer:

Would you describe yourself as a native speaker of English? **YES** ☐ **NO** ☐

Thanks again for making it to the end ☺

Appendix B

Acceptability judgement questionnaire: informants' comments

PAIR	*COMMENTS*
academic training	*job requirement*
acceptable subject	*better: acceptbale topic, maybe?*
active staff	*what are inactive staff??*
actual amount	*tautology*
certified proof	*tautology*
consultative committees	*tautology*
departmental website	*seems strong but i work in a department*
domestic animals	*set phrase*
dynamic region	*unfamiliar*
financial support	*very much in demand*
first year	*association with university schedules*
former graduates	*are they dead?*
francophone country	*see [francophone world]*
french novel	*I studied French at university so for me...*
further information	*see [more information]*
gross salary	*used in advertisements for jobs*
historic buildings	*possible, but not necessarily freq. collocation*
international students	*high number ~ much central funding*
more information	*but with a preposition could be 4/5*
nearest halls	*non specific*
nucleic acids	*technical term*
practical work	*largely tautologous*
real-life scenarios	*uggh!*
recommended gcses	*recommended for what?*
renewable energy	*recenetly become popular in political talk*
responsible investment	*frequent, spreading*
second year	*cf. [first year]*
serious illness	*established collocation*
stand-up comedy	*what else would stand-up go with?*
transferable skills	*management talk - wouldn't say it myself*
video games	*compound?*
web-based systems	*special purpose vocabulary*
work-related learning	*very popular these days*
worth two-thirds	*I don't know what this means without more context*
automated dna	*unfamiliar*
	what does it mean?
coral reefs	*compound?*
	ditto [a fixed phrase]
cystic fibrosis	*too specific, but very frequently used and similar in my L1 (Spanish)*
	a fixed phrase
design-based competition	*unfamiliar*

	i don't know this. assume bureaucracy again
distinguished scholars	*even better: distinguished scientist*
	set phrase
foundation-year entry	*never heard this before*
	I worked in a university with FYE
francophone world	*I can have many different 'worlds'*
	sounds like Italian
fresh insights	*but 'novel' might [be] very strong*
	set phrase in evaluations
key skills	*requirement in job ads*
	management talk again
linear algebra	*sub-discipline*
	a fixed phrase
manual dexterity	*dexterity' implies 'manual'*
	this seems inseparable almost
optional modules	*unfamiliar*
	university curricula
practical skills	*curricula requirements*
	just 'skills' would do
reactive compatibilisers	*unfamiliar*
	can't have a guess what this means
reflexive individuals	*never heard this before*
	this is meaningless to me
strict deadlines	*a nightmare for academic authors*
	high frequency
subatomic particles	*specialist terminology (?) of a different discipline*
	particles' seems like the only word one can have here
sufficient sketchbooks	*what does it mean?*
thermal conversion	*what does it mean?*
	domain specific
unequalled concentration	*unfamiliar*
	style mixture
volcanic eruptions	*very 'trendy'*
	eruptions can be only volcanic
wide range	*ugly*
	very strong!
widest ranges	*not sure it's grammatical*
	singular 'wide(st) range' is stronger
assessed individual	*what does it mean?*
	can't see [unclear writing]
	this one seems different due to the grammatical structure
black holes	*isn't that a compound?*

	might also appear without space
articular cartilage	*unfamiliar*
	medical domain only
	could be strong, i just don't know this term
bioadhesive polymers	*unfamiliar*
	too specific
	probably strong, but not sure
	domain specific
connecting uel	*? unfamiliar*
	what does it mean?
	i don't know this
	? don't know/probably technical term in area i don't know
fast pyrolysis	*unfamiliar*
	what does it mean?
	medical domain
	? don't know/probably technical term in area i don't know
beautiful city	*quite weak, not a collocation*
	possible, but not necessary?
	common collocation
	the word city can be combined with several adjectives
	emotional+basic vocabulary
	frequency
	too individualised
	a city can be evaluated in many ways
	very generic, even travel brochures aim at more descriptive language
	too unspecific
	no problem with this - normal
	just one of many adjs
	seems like something you would often say
	i've been reading guidebooks
	perfectly acceptable, but not mwe
	no comment
	beautiful' is one of many possible adjectives
naked eye	*idiom, strong collocation*
	fixed/fossilized
	it's a common expression in a corpus
	never heard it
	as in not visible to the naked eye
	fully idiomatic: 'to the naked eye'
	idiom

	almost lexicalised
	idiomatic
	set phrase: meaning not derivable from compnents
	idiom - maybe even 5
	there is no other normal succint way of saying 'without optical aid'
	hard to imagine this in a course description
	of course, i'm influenced by Sinclair!
	another fixed phrase
	idiomatic expression (metaphorical)
	as a phraseologist and fan of John Sinclair I would prefer to give this one a '6' (six)
	strongly fixed, idiomatic
open days	*not sure about the collocation*
	compound?
	again, only in academic context
	not very frequently used?
	feels like a compound
	in specialised sense, in institutional domain
	specific referent
	unfamiliar to me, meaning not clear
	if primed in academic context
	doesn't make sense to me
	~opening hours
	possibly idiom-like, uncertain
	a necessary evil
	not one i would recognize
	i would think that 'open day' is more common
	same as above
	compound, effectively - operates as single term
	conceptually clear referrent with very restricted attributive use of 'open'
	very strong in singular in academic context
final year	*don't think of as a strong collocation*
	freq. in academic context
	in academic context; doubt you'll get some results in non-academia
	contrastive to 'last year'
	semantic field closely connected to my professional life
	in an academic context
	in university context only
	i could have a 'second/third' year
	terminology at universities

	if primed in academic context
	related to university life
	association with univeristy schedules
	I perceive this as a compositional phrase
	clear and normal expression
	seems typical of uni catalogs
	this might seem strong because i am a student
	not the strongest
	very common expression
	psychologically salient because of its precise and important referrent
	very strong in academic context
cochlear implants	*technical term, fixed phrase*
	specialised term
	never heard of it
	precision, the term means one particular device
	there are several types of implants
	sort of set phrase (even if you are not native and your L1 is a romance lg.)
	but probably a specialised term
	technical register, only partly everyday English
	technical term: uniq/referent specific
	terminology
	don't know the word
	much propagated in medical context
	I perceive this as a technical term
	don't know any other way of saying this
	? no clue/maybe a technical term
	definitely associated but v. register specific
	does cochlear go with anything else?
	common, effectively a compound
	see above [dramatic text]
	very strong, it's a technical term
	maybe very strong to a particular audience
rigid deadlines	*not sure, sounds unusual*
	(3) 'rigid' more common than 'strict' with 'deadlines'!
	the more i think of it, the stronger degree of lexicalisation I see
	strict would sound more natural
	one could/would express this in diff. ways
	narrow range of evaluative adjs. available a 'deadlines'
	unfamiliar usage, but conveys clear meaning
	a bit Italianate

	correct' collocation
	rather 'strict deadlines' cf. [relevant item]
	syntactically compositional
	strict' seems more associated; 'rigid' by analogy
	not as fixed as 'hard deadlines'
	same as previous [beautiful city]
	one of the typical adjectives used with 'deadline' and one of the typical nouns used with 'rigid'
	i think 'strict' would be stronger than 'rigid'